빅북총서 008

R과 함께하는 통계학의 이해

최용석 부산대학교 자연과학대학 통계학과

BigBook

R과 함께하는

통계학의 이해

R과 함께하는
통계학의 이해

최용석
부산대학교 자연과학대학 통계학과

BigBook

R과 함께하는

통계학의 이해

함께 만들고 함께 나누는 공유의 지식!

인류의 지식은 개인의 것이기에 앞서 문화의 유산입니다. 우리는 물려받은 지식의 토대위에 지식을 창조한 것이며 이는 다음 세대도 그러할 것입니다. 우리의 삶을 풍요롭게 하는 지식은 공기와 같이 공유되어야 하며 이를 통해 더 나은 지식창조가 가능하다고 믿습니다.

이제 지식은 상아탑을 넘어 시민사회의 참여가 필요합니다. 이는 많은 전문가들이 다양한 지식을 가지고 있으며 지식의 변화속도는 상상하기 어려울 정도로 빠르기 때문입니다. 고등교육기관과 시민들이 협력한다면 다양한 견해를 담은 새롭고 혁신적인 지식이 창조될 수 있을 것이며, 함께 나누고 공유한다면 지식은 인류의 삶에 더 큰 기여를 할 수 있을 것입니다.

우선적으로는 교육을 위한 지식들이 공유되어져야 하며 이는 모두에게 평등하게 제공되어야 합니다. 그리하여 문화적인 유산인 지식이 인종과 성별 그리고 지위와 부의 차이에 의하지 아니하고 필요로 하는 모든 사람들에게 다가가 그들에게 보다 나은 삶이 마련되어야 합니다.

고등교육기관의 지식창조 활동 결과물들도 이를 배워야 할 학생들에게 효과적으로 공유될 필요가 있으며, 우리는 이를 위한 노력을 경주할 것입니다. 이제 새롭고 수준 높은 지식을 바라는 우리 이웃들의 목마름을 채우기 위하여 작지만 먼 걸음을 시작합니다.

뜻 있는 많은 분들의 도움으로 먼 길이 외롭지 않기를 바랍니다.

공유와 협력의 교과서만들기 운동본부

머리말 INTRODUCTION

통계학(Statistics)은 불확실하고 잘 알려져 있지 않은 사실과 대상에 대한 통계정보를 얻기 위해 이와 관련된 자료(data)를 수집하고, 그 자료를 요약 정리하여 해석하며, 의사결정을 위한 결론이나 일반성 등을 이끌어내는 데 필요한 이론과 방법을 과학적으로 제시하여 주는 학문이다.

본 교재는 시중에 많은 통계학 입문 수준의 내용과 유사하지만 〔보기〕를 통하여 실제 문제를 이해하고 풀이하는 과정을 보여주는 데 충실히 하였다. 대부분 각 장의 마지막 절의 〈R-프로그램 실습〉은 〔보기〕에서 제시된 통계계산의 편의성을 위해 마련되어 있다.

특히 이 책의 구성과 내용의 간략한 요약은 다음과 같다.

1장 통계학의 이해 우리 주변에서 먼저 통계가 어떻게 활용되는 지를 살펴보고 여론조사나 실험계획에 의한 자료의 수집과 자료의 구성요소인 개체(observation)와 변수(variable)에 대한 이해, 그리고 이를 통한 자료의 종류를 소개한다.

2장 자료의 정리 및 요약 표본으로부터 정보를 획득하기 위해 주어진 자료에 대해 효율적인 방법으로 정리 및 요약하는 기법들을 소개한다. 정리 및 요약의 기법에는 표나 그림을 이용할 수도 있고, 수치적 정보를 이용할 수도 있다.

3장 이산확률변수 및 분포 확률변수의 특징을 이해하기 위해 이산형확률변수를 이용하여 기대값과 분산을 계산하는 방법을 익히고, 더불어 이산확률분포 중에 가장 대표적인 이항분포(binomial distribution)를 활용하는 방법을 익힌다.

4장 연속확률변수 및 분포 연속확률변수와 연속확률변수의 확률분포를 나타내는 확률밀도함수(probability density function)의 특징을 익히고, 더불어 연속확률분포 중에 가장 대표적인 정규분포(normal distribution)를 활용하는 방법을 익힌다.

5장 표집분포와 중심극한정리 표본의 반복추출을 통해 나타나는 표본평균이 가질 수 있는 값들의 특징을 파악하고, 표본평균의 기대값과 분산을 파악한다. 더불어 모집단의 확률분포와는 무관하게 표본의 크기가 충분히 큰 경우 표본평균의 확률분포는 정규분포를 따르게 된다는 중심극한정리에 대해 알아본다.

6장 추정 표본으로부터 획득한 수치적 정보 즉, 통계량을 이용하여 실제 관심의 대상인 모수의 참값에 대해 알아가는 추정의 방법을 익힌다.

7장 가설 검정 : 한 집단의 비교 제기된 주장의 타당성을 검정할 때 통계적 가설 검정의 문제를 다루며 한 집단의 모평균과 모비율을 표본의 크기에 따른 대표본과 소표본에서 다루려 한다. 특히, 기각역을 활용하거나 유의확률 값에 의한 검정방법을 소개한다.

8장 독립표본과 대응표본 : 두 집단의 비교 독립된 두 집단에 의한 독립표본 또는 동일한 한 집단에 대해 두 번 반복 측정한 대응표본을 비교하기 위해선 통계량은 각 집단 또는 두 번 측정한 표본의 평균 또는 비율에 의한 검정을 다루게 된다.

9장 분산분석 : 여러 집단의 비교 여러 집단 또는 여러 처리를 비교할 때 적용되는 분산분석(analysis of variance, ANOVA)에 대해 알아보고 가장 기초가 되는 일원 분산분석에 대해서 논의하고자 한다.

10장 상관분석과 회귀분석 : 두 변수의 관계 여러 분야의 통계분석에서 두 개 혹은 그 이상의 측정변수들의 관계가 중요한 경우가 많다. 이를 측정하기 위한 상관계수(correlation coefficient)와 변수들 간의 관계를 나타내는 함수식을 찾아내고 이를 이용하여 예측 및 추론을 하기 위한 회귀분석(regression analysis)을 소개하고 있다.

11장 분할표 자료분석 : 범주들의 관계 관찰된 자료가 범주형 변수에 따라 정리된 분할표(contingency table)에서 범주간의 독립성 검정(test of independence)과 동질성 검정(test of homogeneity)인 카이제곱 검정을 소개하고 있다.

이 책을 완성하는 데 도움을 준 신상민 박사와 이보희 선생, 그리고 김은성, 박준수, 천선경 대학원 지도학생들의 헌신적인 오류 지적에 고마움을 전합니다. 이 책은 빅북(Big Book) 운동의 일환으로 〈공유와 협력의 교과서 만들기 운동〉 취지에 동참하여 제작되었고 특히, 이런 기회를 제공하신 (사)사회적기업연구원-교과서만들기 운동본부 조영복 대표님의 열정에 존경을 표하며 변은비 연구원의 실무적 도움에도 감사를 드립니다.

2014년 6월

이 책이 모든 이에게 따뜻한 지식공유가 되기를 바랍니다.

저자 드림

목차 CONTENTS

1장

통계학의 이해

R과 함께하는

통계학의 이해

1.1 통계학의 활용

통계학을 이해하기 위해선 먼저 우리 주변에서 통계가 어떻게 활용되는 지를 살펴보는 것이 중요하다. 통계학을 필요로 하는 분야는 농업, 생명과학, 환경과학, 산업연구, 품질관리, 시장조사 등 매우 많다. 우리가 신문을 읽거나, 텔레비전을 보거나, 인터넷을 사용할 때 다음과 같이 수치를 활용한 수많은 통계 정보를 만나게 된다.

■ 실업률

OECD 30개 국가 간 비교를 해보면 한국의 경우 2007년 기준으로 취업률이 63.9%로 OECD 평균인 66.7%보다 낮은 한편 실업률 또한 3.2%로 매우 낮은 수준을 보인다. 이처럼 실업률과 취업률이 함께 가장 낮게 나타난 결과, 일할 의사나 능력이 없는 것으로 간주되는 비경제활동인구 비율이 32.9%로 OECD 평균인 27.7%보다 상당히 높게 나타난다. 또 한국의 장기실업자 비율은 0.6%로 OECD 국가 가운데 가장 낮으며, OECD 평균인 29.1%에 비교할 때 기적 같은 수치다.

■ 농산물 수요 및 소비자 가격

마늘의 경우 5월 이후 지금까지 총 7차례 가격전망을 했는데 단 한 차례만 실제 가격과 소폭 차이가 발생한 것으로 나타났으며, 최근 11월 중순부터 12월 초까지의 예측에서도 도매 가격이 1kg당 2,780원으로 추정범위(2,720~3,000원) 내이다. 특히, 주요 농산물(마늘, 양파, 건 고추, 무, 배추)의 공급과 수요동향, 해외생산 및 수입동향, 재고량, 당월 · 단 · 중기 가격 전망에 대한 정보를 공유하고 있다.

■ 물가 안정

3월 물가 안정은 기상여건 호전, 보육료 지원 확대 등으로 농산물(1.5%→-2.7%)과 개인서비스(0.5%→-0.8%)가 전월 대비 하락세로 전환된 데에 따른 것이라고 통계청과 기획재정부는 설명했다.

■ 지지도 조사

모 후보 지지도가 ‘49%, 2주 만에 반등’한 것으로 조사됐다. 이번 조사는 여론조사 전문기

관 A와 인터넷신문 B와 함께 지난 12월 29~31일 3일간 전국 만 19세 이상 휴대전화가입자 2,500명을 대상으로 임의걸기(RDD) 방식으로 진행했다. 표본은 2013년 11월 말 현재 국가 주민등록인구통계에 따라 비례할당 후 무작위로 추출했고, 표본오차는 95% 신뢰수준에 ±2.0%p다(응답률 : 5.2%).

■ 의학 진단과 예측

당뇨병을 앓은 경우 알츠하이머병에 걸릴 위험이 약 65%까지 높다고 과학자들은 주장했다. 시카고의 러시 의과대학 연구소는 5년 동안 800명을 대상으로 이러한 위험의 증가를 발견했다. 이 시카고의 연구자들은 러시 대학의 종교단에 참가한 824명의 수녀와 신부 그리고 수사들을 대상으로 5년 반 동안 연구하였다. 이 연구과정에서 151명의 사람들이 알츠하이머에 걸렸고 그 중에서 31명은 당뇨병 환자들이었다. 연구자들은 계산결과 당뇨병 환자들이 비당뇨병 환자들에 비해 알츠하이머에 걸릴 위험이 65%나 높다는 사실을 밝혀냈다.

■ 사회경제적 지위와 저체중아 연관성

통계청의 1995년부터 2004년까지 전국의 출생신고 자료에 등록된 출산아 580만 1,545명을 분석한 결과 초등학교 졸업 학력의 산모는 대졸 이상인 산모보다 저체중아를 낳을 가능성이 1.77배나 높은 것으로 나타났다. 중졸 산모도 대졸 이상보다 저체중아 출산 가능성이 1.46배 정도 높았다. 부모의 낮은 사회 · 경제적 지위가 자녀의 나쁜 건강상태로 이어진 것이다. 저체중아는 태어날 때 몸무게가 2.5㎏ 미만인 아기를 가리킨다. 아버지의 학력도 저체중아 출생과 관련이 있었는데, 초등학교 졸업 이하는 대졸 이상보다 저체중아를 얻을 확률이 1.69배 높았으며, 고졸 아버지는 1.1배, 중졸은 1.44배 각각 높았다.

■ 시청률 조사

새해 첫 날인 어제(2일) 방송된 드라마 중 시청률 1위는 KBS1 일일드라마 ‘사랑은 노래를 타고’가 차지했다. 3일 시청률 조사기관에 따르면 전날 방송된 드라마 중 ‘사랑은 노래를 타고’는 27.7%로 1위에 올랐다. 전날 26.9%보다 0.8%로 상승한 수치다. 전체 드라마 시청률 2위이자, 수목드라마 시청률 1위를 기록 중인 SBS ‘별에서 온 그대’는 24.6%로 집계됐다. 앞서 방송분보다 2.3% 상승했다. KBS2 ‘루비반지’는 22.5%(3위)로, KBS2 ‘TV소설 은희’는 16.9%(4위)로 조사됐다. 뒤이어 SBS ‘두 여자의 방’(14.7%)이 5위에 랭크됐다. 수목드라마 MBC ‘미스코리아’는 8.9%로 전날 방송(9.5%)보다 다소 하락했다.

■ 교통진단

빅데이터 분석, 처리 기술을 교통 분야에 접목하였다. 부산시 과거 각종 교통정보를 활용하여 교통 정체구간을 시공간적 길이로 분류 패턴화하였고 주변 도로의 교통량 등에 영향을 미치는 변수를 추가로 분석 결합해서 현재 이후의 교통 흐름을 구간별, 시간대별로 예측하였다. 앞으로 날씨와 사고 여부 정보까지 반영한 교통 빅데이터 예측 정확도 향상 연구를 추진할 계획이다.

이상에서 살펴본 바와 같이 수치를 활용한 통계정보는 사회경제적 지위와 저체중아 연관성과 같이 현재 관찰되는 사회적 현상일 수도 있고, 물가안정을 가져다 준 요인을 찾아내기도 하며, 농산물 수요와 가격을 예측하기도 하여 우리들이 의문점을 갖는 다양한 분야에 대한 해답을 제시한다. 더불어 이러한 해답이 기존 사실과 비교해 얼마나 위배되며 어느 정도 믿을 수 있는지 확인해보는 것도 중요하다. 우리는 믿을 수 있는 정보를 통하여 합리적이고 과학적인 방향으로 의사결정을 내리게 된다. 최근 교통진단과 같이 사회 여러 분야에서 축적된 대용량 자료로부터 정보를 추출하고 이를 토대로 예측하거나 대응하기 위한 빅테이터 기술에 대한 국가적 관심이 높다.

따라서 불확실하고 잘 알려져 있지 않은 사실과 대상에 대한 통계정보를 얻기 위해 이와 관련된 자료(data)를 수집하고, 그 자료를 요약 정리하여 해석하며, 의사결정을 위한 결론이나 일반성 등을 이끌어내는 데 필요한 이론과 방법을 과학적으로 제시하여 주는 학문으로 통계학을 이해하면 될 것 같다.

통계학의 정의와 분류

통 계 학 : 불확실하고 잘 알려져 있지 않은 사실과 대상에 대한 통계정보를 얻기 위해 이와 관련된 자료(data)를 수집하고, 그 자료를 요약 정리하여 해석하며, 의사결정을 위한 결론이나 일반성 등을 이끌어내는 데 필요한 이론과 방법을 과학적으로 제시하여 주는 학문이다.

기술통계학 : 자료를 정리하여 그림이나 표로 요약하거나 자료들의 수치값을 요약한 대표값이나 자료의 흩어진 형태(분포)와 변동의 크기 등을 구하는 분야.

추측통계학 : 통계적 모형과 구간을 설정하기도 하고 추측하기도 하며 어떤 기존의 사실에 대하여 가설을 세우고 이를 검정하고 예측하는 분야.

통계학에는 자료를 수집하고, 그 자료를 정리하여 그림이나 표로 요약하거나 자료들의 수치값을 요약한 실업률, 소비자물가 지수, 시청률, 소득평균 등과 같이 대표값이나 자료의 흩어진 형태(분포)와 변동의 크기 등을 구하는 방법을 다루는 기술통계학(descriptive statistics)이 있다. 이를 통하여 자료의 전반적인 형태와 특징을 파악할 수 있게 된다. 실제 통계학이 불확실하고 잘 알려져 있지 않은 사실을 발견해내고 미래에도 이런 사실이 지속될 것인지를 예측하고 이들의 믿을 만한 정도를 평가하는 데 큰 위력을 발휘한다고 생각한다. 이를 위해서 일반적으로 통계적 모형과 구간을 설정하기도 하고 추측하며 어떤 기존의 사실에 대하여 가설(hypothesis)을 세우고 이를 검정(test)하는 추측통계학(inference statistics)이 현대 통계학에서 핵심이 되고 있다.

1.2 자료의 수집

기술통계학과 추측통계학을 적용하기 위한 통계정보를 얻기 위해선 자료를 수집해야 하는데 사회 현상의 정보를 얻기 위한 사회과학 분야에서는 표집조사(sampling)와 여론조사(survey) 방법을 사용하고, 모의실험이나 측정을 통한 정보를 얻는 자연과학 분야에서는 실험계획(experimental design) 방법을 주로 사용한다. 예를 들자면 대통령 후보에 대한 자질도나 호감도를 조사하기 위해선 지역, 나이, 교육수준 등을 고려한 적절한 표집방법을 잘 선택하여 전체를 잘 대표할 사람들을 대상으로 여론조사를 실시해야 한다. 그리고 환경의 변화에 따른 혈압과 수명을 알아보기 위해선 동일한 대상에 대하여 환경변화 전과 후를 잘 계획하여 측정해야 한다.

사회과학 분야나 자연과학 분야나 통계정보를 얻기 위한 자료의 수집에는 두 가지 방법이 있다. 관심 있는 대상의 전체를 조사하는 전수조사가 있고 일부만 조사하는 표본조사가 있다. 이는 조사시간과 비용을 고려하여 이루어진다. 인기 드라마에 대한 시청률 조사인 경우 전 국민을 대상으로 조사하기란 불가능하다. 이럴 경우 적절한 표본의 패널을 선정하고 이들을 통하여 드라마 시청을 한 사람을 집계하게 된다. 반면에 전수조사인 총인구조사(census)인 경우 전체 대상의 수가 너무 많아 조사시간이 오래 걸리며 조사원들의 교육과 관리의 집중화와 효율성도 매우 중요한 요소이다. 엄청난 제작비용을 지불해야 하는 인공위성이나 로켓의 성능 실험을 위해 전수조사를 한다는 것은 불가능한 일이다. 이와 같이 전체 자료의 수가 적은 경우를 제외하면 표본조사를 통해 통계조사가 이루어지는 것이 바람직하다.

일반적으로 통계조사의 관심이 되는 모든 대상들의 집합을 모집단(population)이라고 부른다. 특히, 모집단의 특성을 나타내는 수치를 모수(parameter)라 한다. 시청률 조사에서 전 국민을 모집단으로 하는 시청률이 모수에 해당한다. 표본(sample)은 시청률 조사를 위해 선정된 패널에 해당하며 이 표본의 특성을 나타내는 수치를 통계량(statistics)이라 하며 표본을 통해 얻어진 시청률이 여기에 해당한다. 모집단과 표본의 이해를 위해 다음의 보기를 살펴보자.

보기 1.1

부산시가 시 전체 노동력 조사를 대신하여 남구와 서구의 각 5개 구별 대상으로 노동력 인구와 이들 중 실업자의 비율을 알고자 한다.

- 모집단 : 부산시 전체 노동력 인구와 실업자
- 모　수 : 모집단의 노동력 있는 인구 수, 실업자 수, 실업자의 비율
- 표　본 : 남구와 서구의 각 5개 구별 노동력 인구와 실업자
- 통계량 : 표본의 노동력 있는 인구 수, 실업자 수, 실업자의 비율

보기 1.2

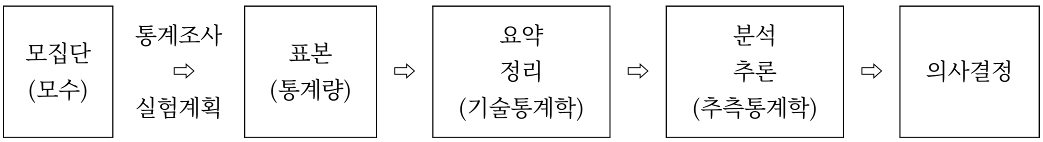
A 타이어 공장의 5개 공정과정에서 하루에 생산되는 사계절용 타이어 10,000개의 평균 주행거리를 알고자 실험을 실시하였다. 실제로 각 공정과정에서 10개씩 50개 타이어를 대상으로 평균 주행거리를 계산하였다.

- 모집단 : A 타이어 공장의 5개 공정과정에서 하루에 생산되는 사계절용 타이어 10,000개
- 모　수 : 모집단의 타이어의 평균 주행거리
- 표　본 : 각 공정과정에서 10개씩 50개 타이어
- 통계량 : 표본의 타이어의 평균 주행거리

[보기 1.1]과 [보기 1.2]를 통하여 1.1절에서 언급한 추측통계학에 대한 이해를 추가적으로 설명하자. 각각의 보기를 통하여 알고자 하는 통계정보는 부산시 노동력 인구와 이들 중 실업자의 비율 또는 부산시 전체 노동력 인구와 실업자 또는 A 타이어 공장의 5개 공정과정에서 하루에 생산되는 사계절용 타이어 10,000개의 평균 주행거리이다. 이 통계정보를 모수라고 볼 수 있다. 이를 전수조사를 통하여 비용이나 시간적 제약으로 얻기 어렵기 때문에 모집단으로부터 추출된 표본을 통한 통계량에 의해서 추측하게 된다. 이렇게 모수를 추측함으로써 얻어낸 통계량의 신뢰성을 살펴보는 것 또한 당연히 이루어져야 할 과제이다.

[그림 1.1]은 지금까지 살펴본 통계학의 정의와 통계정보를 얻기 위한 자료수집 등 통계자료분석의 단계를 한눈에 살펴보도록 그림으로 요약한 것이다.

[그림 1.1] 통계자료분석의 단계



1.3 자료의 이해

자료의 수집을 위해 제일 먼저 이해되어야 할 부분은 자료의 구성요소인 개체(observation)와 변수(variable)이다. 변수는 자료의 수집에서 관심이 되며 측정해야 할 어떤 특성을 말하며 설문지 각 문항의 응답, 성별, 키, 몸무게, 온도, 혈압, 머리 색깔, 과목 성적, 철의 강도, 주행시간, 단백질량, 실업자 수, 사망자 수, 사고건수, 주행거리, 사고의 정도 등을 들 수 있다. 이들에 의해서 실제 측정되는 대상으로 가장 기본이 되는 단위를 개체라 하며 50명의 남녀, 30마리의 쥐, 5개 공정과정에서 하루에 생산되는 사계절용 타이어 10,000개, 80명의 학생 등이 있다. 대개 변수에 의한 자료의 측정값은 변수값이 된다.

[표 1.1]의 자료를 살펴보면 10명의 고등학생을 대상으로 흡연, 성별, 키, 몸무게, 평소 운동량을 측정하였고 이들 중 달리기를 한 사람과 달리기를 하지 않은 사람으로 나누고 이들의 처음 심장박동수와 달리기를 마친 후 모든 학생의 나중 심장박동수를 측정하였다. 여기서 10명의 고등학생이 개체가 되고, 심장박동수, 흡연, 성별, 키, 몸무게, 평소 운동량을 변수라 하며 이들은 다음과 같이 변수값을 가진다.

달리기(1=달리기를 한 사람, 2= 달리기를 하지 않은 사람), 흡연(1=흡연자, 2=비흡연자), 성별(1=남자, 2=여자), 운동량(1=적음, 2=보통, 3=많음).

[표 1.1] 심장박동수 자료

학생	처음 심장 박동수	나중 심장 박동수	달리기	흡연	성별	키 (cm)	몸무게 (kg)	운동량
1	64	88	1	2	1	168	63.5	2
2	58	70	1	2	1	183	65.8	2
3	62	76	1	1	1	185	72.6	3
4	66	78	1	1	1	185	86.2	1
5	64	80	1	2	1	175	70.3	2
6	64	60	2	2	2	168	81.6	3
7	94	92	2	1	2	157	82.1	2
8	60	66	2	2	2	157	54.4	2
9	72	70	2	2	2	173	78.5	2
10	58	56	2	2	2	170	56.7	2

자료는 측정변수의 성질에 따라 일반적으로 질적 자료(qualitative data)와 양적 자료(quantitative data)로 나눌 수 있고 이들 두 가지가 혼합된 경우가 대부분이다. [표 1.1]의 경우 질적 자료는 변수가 달리기(1=달리기를 한 사람, 2=달리기를 하지 않은 사람), 흡연(1=흡연자, 2=비흡연자), 성별(1=남자, 2=여자), 운동량(1=적음, 2=보통, 3=많음)에 의해서 측정된 자료이다. 즉, 질적이란 개체인 측정대상이 어느 범주에 들어가는지를 나타내며 이를 범주형 자료(categorical data)라고도 한다. 양적 자료는 변수가 처음 심장박동수, 나중 심장박동수, 키(cm), 몸무게(kg)와 같이 양을 나타내는 수치자료를 말한다.

특히, 양적 자료에서 키와 몸무게와 같이 관측값들이 셀 수 없는 연속적인 구간의 값을 취하는 경우를 연속형 자료(continuous data)라 하고, 심장박동수와 같이 관측값들이 셀 수 있는 값을 취하는 경우를 이산형 자료(discrete data)라 한다. 질적 자료에서 운동량의 경우 세 가지 범주(1=적음, 2=보통, 3=많음)가 주어지며 이들 사이는 순서상 의미가 있으므로 순위형 자료(odinal data) 또는 서수척도 자료(ordinal scaling data)라 한다. 반면에 달리기(1=달리기를 한 사람, 2=달리기를 하지 않은 사람), 흡연(1=흡연자, 2=비흡연자), 성별(1=남자, 2=여자)과 같은 질적 자료는 수치가 단지 구분하기 위한 것으로 이를 명목형 자료(nominal data) 또는 명목척도 자료(nominal scaling data)라 한다.

양적 자료와 질적 자료는 서로 변환이 가능하지만 질적 자료에서 양적 자료로의 변환은 다소 고급 개념이 요구되므로 생략하기로 하자. 자료를 요약하는 초기 단계에서 양적 자료의 경우 수치값을 이용하므로 평균과 분산과 같은 통계적 계산을 위해 사용되지만 질적 자료의 경우는 범주에 따른 자료의 빈도수를 이용하므로 자료를 요약 정리하는 데 용이하다.

보기 1.3

[표 1.1]로부터 양적 자료인 키의 자료를 150에서 170 사이를 1=하, 171에서 180 사이를 2=중, 181 이상을 3=대로 변환하면 다음의 질적 자료를 얻는다.

양적 자료:	168	183	185	185	175	168	157	157	173	170
질적 자료:	1	3	3	3	2	1	1	1	2	1

따라서 자료는 조사방법과 측정변수에 따라 다양하게 구분할 수 있지만 크게는 양적 자료와 질적 자료로 구분되고 양적 자료에는 연속형과 이산형으로 나누고, 질적 자료에는 순위형과 명목형으로 나눌 수 있음을 살펴보았다.

■ 자료의 종류

- 자료
  - 양적 자료
    - 연속형 자료(예 : 키)
    - 이산형 자료(예 : 입장인원 수)
  - 질적 자료
    - 순위형 자료(예 : 선호도)
    - 명목형 자료(예 : 성별)

1.4 연습문제

1.1 다음의 자료를 양적 자료와 질적 자료로 분류하여라.

(1) 판매된 콘서트 티켓 수
(2) 체지방 양
(3) 좋아하는 야구 팀
(4) 식료품을 구입하기 위해 식료품 코너에 서 있는 시간
(5) 부산대학교에 등록된 학생 수
(6) 가장 많이 본 TV 프로그램
(7) 치약의 브랜드
(8) 가장 가까운 극장까지의 거리
(9) 국내 10대 기업 임원의 나이

1.2 A 피트니스 센터는 회원들이 일주일에 운동하는 시간의 평균을 조사하기 위하여 전체 회원들 중 50명을 뽑아서 조사하였다. 이 경우 다음을 정의하라.

(1) 모집단
(2) 모수
(3) 표본
(4) 통계량

1.3 결혼 상담사는 결혼을 지속하는 고객들의 비율을 알기 위하여 자신이 상담한 고객들 중 100명을 뽑아서 조사하였다. 이 경우 다음을 정의하라.

(1) 모집단
(2) 모수
(3) 표본
(4) 통계량

1.4 다음의 자료를 명목척도자료, 서수척도자료, 구간척도자료, 비율척도자료로 분류하여라.

(1) 레스토랑 메인 요리들의 베이킹 온도

(2) 24색 크레파스 세트 안에 든 크레파스 색깔

(3) 대형마트의 월 소득

(4) 웹사이트 만족도 조사: 매우 만족, 보통, 매우 불만족

(5) 정치적 견해: 극좌파, 좌파, 우파, 극우파

(6) 아날로그 시계의 눈금

(7) 21-65세 여성의 키

(8) 통계학 과목 수강생들의 평점: A, B, C, D, F

1.5 한 통계학과 교수는 통계학 수강생들의 한 학기 동안 평균 결석일 수에 관심이 있다. 통계학 수강생들 중 10명을 뽑아서 조사한 결과, 평균 결석일 수는 3.5일로 나타났다.

(1) 모집단은 무엇인가?

(2) 표본은 무엇인가?

(3) 통계량은 얼마인가?

1.6 대구에 위치한 지역 공원을 이용하는 주민들의 나이, 주당 이용횟수, 이용시간을 알기 위하여 공원 주변의 집들 중 일부를 선택하여 조사하였다. 다음의 자료를 연속형과 이산형으로 분류하여라.

(1) 주민들의 나이

(2) 주당 이용횟수

(3) 이용시간

1.7 부산 금정구 보건소에 방문한 성인 20명의 키와 몸무게를 조사하여 체질량지수(BMI)를 계산하였다. 체질량지수가 20 미만은 저체중, 20 이상~25 미만은 정상체중, 25 이상~30 미만은 과체중, 30 이상~35 미만은 비만, 35 이상은 고도비만으로 자료를 변환시켰다. 변환하기 전과 후의 자료 형태는 각각 무엇인가?

2장

자료의 정리 및 요약

R과 함께하는

통계학의 이해

2.1 범주형 자료의 요약

일반적으로 범주형 자료에서는 몇 개의 범주가 반복해서 나타나게 되므로, 각 범주가 나타나는 횟수를 요약함으로써 범주형 자료의 내용을 쉽게 파악할 수 있을 것이다. 이처럼 이 절에서는 범주형 자료의 내용을 효과적으로 파악하기 위해 표나 그래프를 이용하여 자료를 정리/요약하는 기법을 공부한다.

우선, 범주형 자료의 요약에 대해 알아보자. 범주형 자료가 가지는 각각의 관측값은 몇 개의 범주 중 하나의 값을 가지게 되므로, 범주형 자료를 요약할 때는 각각의 범주에 속하는 관측값의 개수를 파악하게 된다. 이를 도수(frequency)라고 한다. 범주형 자료에 대해 각각의 범주와 그에 대응되는 도수를 나열한 표를 도수분포표(frequency table)라고 하며, 이러한 도수분포표를 작성하는 것은 범주형 자료에 대한 가장 기본적인 요약기법이다. 도수분포표에는 필요에 따라 추가적인 정보를 나타내기도 하는데, 도수를 자료의 전체 개수로 나눈 비율인 상대도수(relative frequency)를 추가적으로 기록한다. 이러한 상대도수는 때때로 백분율(%)로 표현하기도 한다.

보기 2.1

어느 대학에서 통계학 수업을 수강하는 55명의 학생들을 대상으로 혈액형을 조사한 결과는 다음과 같다. 이 자료를 도수분포표로 요약하라.

B	A	B	A	A	B	O	A	A	A	O
B	AB	B	AB	AB	A	A	O	AB	O	A
B	O	B	B	A	A	O	A	A	AB	B
B	O	B	B	B	A	AB	A	A	B	O
B	B	O	B	O	B	A	A	AB	A	A

혈액형이 A인 학생의 수 즉, 도수는 20이며 이에 대한 상대도수는 (20/55=) 0.364이다. 이와 같이 각각의 혈액형별로 도수와 상대도수를 파악하고 도수분포표를 나타내면 다음과 같다.

혈액형	도수	상대도수
A	20	0.364
B	18	0.327
O	10	0.182
AB	7	0.127
합계	55	1.000

어떠한 자료든지 해당 자료를 요약해서 나타낼 때는, 단순한 숫자나 문자의 나열보다는 그림을 이용하는 것이 전반적인 개요를 쉽게 이해할 수 있다. 범주형 자료를 요약할 때 사용할 수 있는 효과적인 그래프로는 원도표(pie chart)와 막대도표(bar chart)가 있다.

원도표는 원을 그린 후, 상대도수에 비례하게 중심각을 나누어 해당 원을 조각내서 표현한 그림이다. 원도표를 이용하면 각 범주 또는 몇 개의 범주가 차지하는 비율을 파악하기 용이해진다. 그러나 범주상의 도수를 비교하거나 도수의 차이를 파악하기에는 힘든 경우가 많다. 이러한 경우 이용할 수 있는 그림이 막대도표이다.

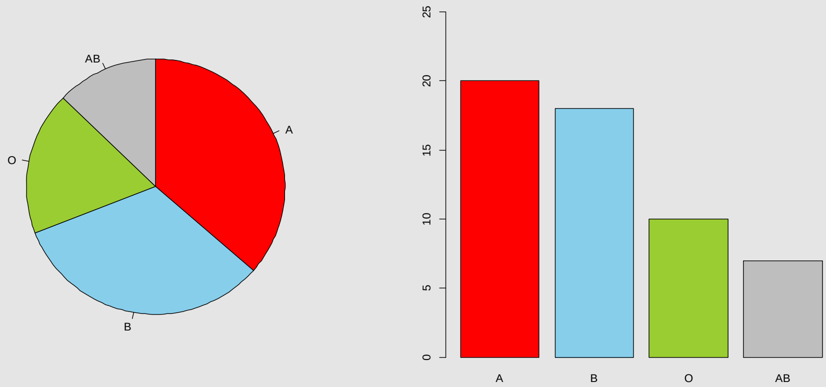
막대도표는 각 범주에서 도수의 크기를 막대의 높이로 표현한다. 만약 전체에서 각 범주가 차지하는 비율에 관심이 있으면 도수 대신 상대도수를 사용하기도 한다. 도수를 이용하든 상대도수를 이용하든 그림의 모양은 같다. 이러한 막대도표를 이용하면 각 범주간의 도수를 비교하는데 용이하다.

보기 2.2

[보기 2.1]에 제시되어 있는 55명의 혈액형 자료에 대한 원도표와 막대도표로 요약하라.

원도표를 그리기 위해서는 표현할 원조각의 중심각을 계산하여야 한다. A형의 경우 상대도수가 0.364이므로 360°이므로 A형을 나타낼 조각의 중심각은 약 (360°× 0.364 ≒)131°이다. 이와 같이 계산하면 B형은 118°이고 O형은 65°, AB형은 46°가 된다. 이를 이용하여 원도표를 그리면 [그림 2.1]의 (a)와 같다. 막대도표는 막대의 높이가 해당 범주의 도수를 나타내며, 그 결과는 [그림 2.1]의 (b)와 같다.

[그림 2.1]의 (a) 원도표를 보면 55명의 학생 중 혈액형이 A형과 B형인 학생의 비율이 높음을 알 수 있다. 그러나 둘 중 어느 혈액형이 더 많은지는 구분하기 힘들다. 이런 경우, [그림 2.1]의 (b) 막대도표를 이용하면 A형인 학생의 수가 B형인 학생의 수보다 많음을 쉽게 알 수 있다.



(a) 원도표 (b) 막대도표

[그림 2.1] 혈액형 자료에 대한 원도표와 막대도표

2.2 이산형 자료의 요약

앞서 언급한 바와 같이 이산형 자료(discrete data)는 관측값을 측정할 때 세어서 파악한 자료를 말한다. 이러한 이산형 자료는 관측값 중 중복되는 값이 많고 적음에 따라 요약하는 기법이 다르다. 일반적으로 중복되는 값이 많으면 2.1절에서 소개한 범주형 자료의 요약기법을 사용하고, 중복되는 값이 적으며 다음 절에서 소개하게 될 연속형 자료(continuous data)의 요약기법을 사용한다.

예를 들어, 특정 도시에서 각 가정의 자녀수를 조사한다고 가정하자. 이러한 조사의 관측값들은 대부분 0, 1, 2, 3, 4 등의 값들로 기록될 것이다. 다시 말해 0에서 4 사이의 값들이 많이 중복될 것이다. 이러한 경우에는 0에서 4 사이의 각각의 값들을 하나의 범주로 간주하여 도수분포표를 작성하고 막대도표나 원도표를 그려서 요약할 수 있다.

그러나 특정 도시의 야구장에 매 경기 입장한 관객의 수를 조사한다고 가정하면, 이러한 조사의 관측값들은 거의 대다수가 서로 다른 값을 가지게 될 것이다. 이러한 경우 범주형 자료의 요약기법에 따라 도수분포표를 작성하게 되면 중복되는 값이 없기 때문에 모든 도수가 1인 도수분포표가 만들어지게 된다. 이와 같은 경우에는 해당 자료를 연속형 자료로 간주하여 다음 절에서 소개되는 연속형 자료의 요약기법을 사용할 수 있다.

2.3 표와 그림을 이용한 연속형 자료의 요약

연속형 자료는 연속적인 척도에 의해 관측되는 자료로 반올림되어 정수값으로 표현되기도 하지만, 이산형 자료와는 달리 실제값은 실수값으로 표현될 수 있다. 이와 같은 이유로, 연속형 자료는 관측값들 중에서 중복되는 값이 많지 않을 수도 있다. 이러한 연속형 자료를 요약할 때에는 여타 자료와는 달리 좀 더 세밀한 주의가 필요하다.

도수분포표는 기본적으로 범주형 자료의 요약기법이다. 따라서 관측값들 중에서 중복되는 값이 많은 경우 즉, 관측값의 종류가 적은 경우에 유용하게 사용할 수 있다. 그러나 연속형 자료는 일반적으로 관측값의 종류가 많기 때문에 최소값부터 최대값까지 모든 관측값을 포함하는 범위를 몇 개의 구간으로 나누어 각 구간에 포함되는 관측값의 개수를 도수로 표현하는 방식으로 도수분포표를 작성하게 된다. 이때, 나뉘어진 각 부분을 계급(class)이라 하고 각 계급에 포함되는 값의 범위를 계급구간(class interval)이라 한다.

연속형 자료를 위한 도수분포표의 계급구간 설정방법을 구체적으로 설명하자면, 우선 주어진 자료의 최대값과 최소값의 차이를 계산하여 모든 관측값을 포함하는 범위를 파악한 후, 계급의 개수로 나누어 계급구간의 폭을 결정한다. 이때, 계급의 개수를 결정할 때에는 특별한 법칙이 있는 것은 아니다. 다만, 계급의 개수가 너무 적으면 계급구간의 폭이 넓어져서 해당 구간의 도수가 커지게 되므로 너무 많은 정보를 잃어버리게 된다. 반대로 계급의 개수가 너무 많으면 계급구간의 폭이 좁아질 것이다. 이러한 경우에는 도수가 나타내게 되는 어떠한 경향도 찾기 힘들어지게 된다. 따라서 계급의 개수를 정할 때는 자료의 성향을 파악하여 도수의 분포 경향이 잘 드러날 수 있도록 계급의 개수를 선택하여야 한다. 이와 같은 이유로 연속형 자료를 도수분포표로 표현할 때에는 많은 시행착오를 겪게 된다.

다음으로 계급의 폭에 따라 모든 관측값을 포함하도록 계급구간의 경계점을 설정한다. 이 경우, 관측값이 계급의 경계점에 놓이지 않도록 하는 것이 바람직하기 때문에 계급구간의 경계점 값은 관측 단위보다 한 단계 아래의 단위로 잡기도 한다. 또한 '첫 번째 계급구간의 시작점과 최소값까지의 거리'와 '마지막 구간의 종료점과 최대값까지의 거리'가 비슷하게끔 조절해 주는 것이 좋다. 그 이유는 첫 번째 계급구간의 시작점에 최소값이 위치하게 되면 마지막 계

급구간에 최대값이 포함되지 못하고 한 단계 앞 의 계급구간에 최대값이 위치할 수도 있기 때문이다.

보기 2.3

다음의 자료는 어느 대학에서 임의로 선정한 남학생 55명의 키를 기록한 것으로 단위는 센티미터(㎝)이다. 이 자료에 대한 도수분포표를 작성하라.

170	178	171	168	173	178	171	174	170	170	175
170	169	166	162	170	171	175	175	171	171	170
172	179	164	170	181	178	180	177	166	169	168
165	163	175	166	178	165	168	167	177	168	177
174	174	176	179	169	173	167	170	173	170	162

우선 주어진 자료의 최대값과 최소값의 차이는 $181 - 162 = 19$이다. 이 자료에 대해 도수분포표를 만들기 위해 계급의 개수를 5개로 결정하게 되면 계급구간의 폭은 $19/5 = 3.8$이다. 이와 같은 경우 계급구간의 폭은 해당 값을 올림하여 4로 하는 것이 좋다.
다음으로 첫 번째 계급구간의 시작점과 최소값까지의 거리와 마지막 구간의 종료점과 최대값까지의 거리가 비슷하게끔 조절하면, 첫 번째 계급구간의 시작점을 161로 설정할 수 있다. 이 경우 계급구간의 경계점 값이 관측 단위와 같아지게 되므로 좋지 않다. 즉, 계급구간의 경계점 값은 관측 단위보다 한 단계 아래의 단위가 될 수 있게끔 첫 번째 계급구간의 시작점을 161.5로 시작하는 것이 좋다. 이와 같은 방법으로 계급구간을 설정하고 도수분포표로 나타내면 다음과 같다.

계급구간(㎝)	도수	상대도수
161.5 이상 165.5 미만	6	0.109
165.5 이상 169.5 미만	12	0.218
169.5 이상 173.5 미만	18	0.327
173.5 이상 177.5 미만	11	0.200
177.5 이상 181.5 미만	8	0.146
합계	55	1.000

연속형 자료에서도 도수분포표를 기초로 하여 각 계급에 대하여 범주형 자료에서의 막대도표와 유사한 모양의 그림을 그릴 수 있는데, 이러한 그림을 히스토그램(histogram)이라 부른다. 막대도표는 막대의 높이가 도수 혹은 상대도수를 나타내어 막대의 높이만으로 각 범주 간의 도수를 비교할 수 있으나, 히스토그램을 이용하면 연속형 자료에 대해 각 계급구간 사이의 도수를 비교하는 것 외에도 계급구간에 따른 도수 변화의 경향 즉, 자료의 분포를 쉽게 알아낼 수 있다.

히스토그램은 막대도표와는 달리 막대의 넓이가 상대도수를 나타내게 된다. 따라서 히스토그램의 전체 면적은 항상 1이 된다. 히스토그램에서 막대의 높이는 상대도수를 계급구간의 폭으로 나눔으로써 구할 수 있는데, 만약 계급구간의 폭이 모두 동일한 경우에는 상대도수를 막대의 높이로 사용하여 히스토그램을 그리더라도 같은 모양으로 표현된다. 그러나 계급구간의 폭이 일정하지 않은 경우에는 서로 다른 모양으로 표현되므로 주의를 기울여야 한다.

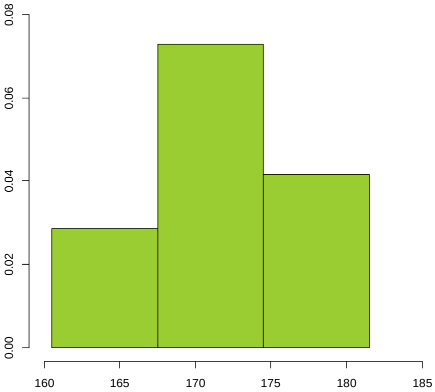
보기 2.4

[보기 2.3]의 55명 남학생들의 키 자료에 대한 히스토그램을 그려라.

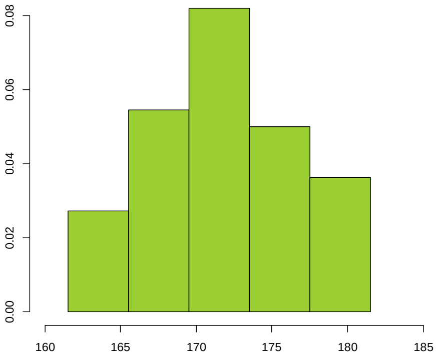
히스토그램을 그리기 위해서는 막대의 높이를 계산하여야 한다. 따라서 [보기 2.3]에서 작성한 도수분포표는 계급구간의 폭이 4이므로, 이를 이용하여 각각의 계급에 해당하는 막대의 높이를 계산하면 다음과 같다.

계급구간(㎝)	상대도수	높이
161.5 이상 165.5 미만	0.109	0.027
165.5 이상 169.5 미만	0.218	0.055
169.5 이상 173.5 미만	0.327	0.082
173.5 이상 177.5 미만	0.200	0.050
177.5 이상 181.5 미만	0.146	0.036

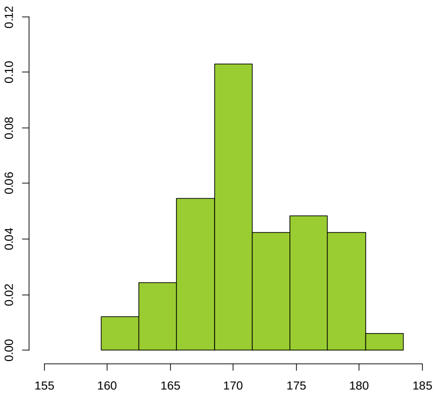
따라서 히스토그램은 다음과 같이 표현된다.



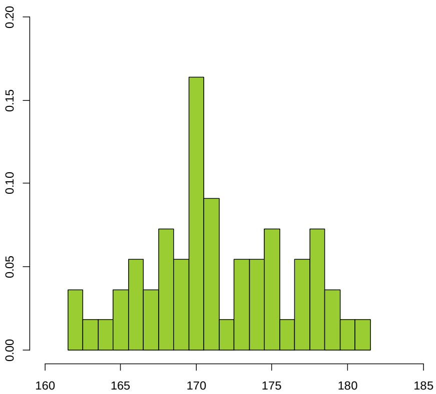
(a) 계급의 수 : 3



(b) 계급의 수 : 5



(c) 계급의 수 : 8



(d) 계급의 수 : 20

[그림 2.3] 계급의 개수에 따른 히스토그램의 변화

도수분포표의 작성과정에서 계급의 개수를 결정할 때, 계급의 개수가 너무 적으면 너무 많은 정보를 잃어버리게 되고, 반대로 계급의 개수가 너무 많으면 도수가 나타내게 되는 어떠한 경향을 찾기 힘들어지게 된다는 것을 언급한 바 있다. 따라서 계급의 개수를 정할 때는 자료의 성향을 파악하여 도수의 분포 경향이 잘 드러날 수 있도록 계급의 개수를 선택하여야 하는데, 도수분포표 상에서의 계급의 개수는 히스토그램 상에서의 막대의 개수와 일치한다. 히스토그램의 모양을 생각하면서 계급구간의 폭과 시작값을 변화시켜가며 계급의 개수를 결정하

는 것이 좋다. 이러한 작업은 수차례의 시행착오를 겪어야 하기에 수작업으로 진행하기는 어려우므로 통계 소프트웨어를 사용하여 여러 히스토그램을 비교하게 된다.

[그림 2.3]은 [보기 2.3]의 55명 남학생들의 키 자료에 대해 계급의 개수 즉, 막대의 개수에 따른 히스토그램의 모양 변화를 나타내고 있다. 그림을 통해 확인할 수 있듯이 계급의 개수가 너무 적으면 특정 계급구간에만 도수가 몰려 전반적인 분포의 성향을 파악하기 힘들고, 반대로 계급의 개수가 너무 많으면 히스토그램의 막대들이 들쑥날쑥해짐으로서 이 또한 전반적인 분포의 성향을 파악하기 힘들다. 따라서 [그림 2.3]의 그림 중에서는 계급의 수를 5개로 하는 경우가 분포의 모양을 잘 드러내고 있다고 할 수 있다.

2.4 수치를 이용한 연속형 자료의 요약

앞 절에서는 도표나 그림을 이용하여 자료를 요약함으로서 관측값들의 분포상태를 시각적으로 쉽고 빠르게 이해할 수 있는 방법을 익혔다. 그러나 이러한 방법들은 일관성과 객관성이 부족할 뿐만 아니라, 통계적 추론에서 요구되는 이론적 근거를 제시하는 것이 쉽지 않다. 따라서 이 절에서는 연속형 자료가 어떤 값을 중심으로 얼마만큼 흩어져 있는가를 나타내는 구체적인 수치정보로 요약하는 방법을 다루고자 한다.

표본평균은 중심위치의 측도 중에서 가장 많이 사용되는 방법으로 관측값의 총합을 관측값의 개수로 나눈 것이다. 이러한 표본평균을 계산하는 방법은 다음과 같다.

■ 표본평균(sample mean)

n 개의 표본자료를 $x_1, x_2, \cdots, x_n$ 이라 할 때, 표본평균 $\overline{x}$ 는 다음과 같이 계산한다.

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

이러한 표본평균은 6장 이후부터 다루게 되는 통계적 추론과정에서 광범위하게 사용되며, 통계적 분석과정에서 가장 기초적인 통계 수치이다. 그러나 표본평균은 모든 관측값이 반영되므로, 극단적으로 아주 크거나 작은 값에 영향을 많이 받아 때로는 잘못된 중심위치를 나타내기도 한다. 이러한 경우 사용할 수 있는 것이 중위수이다.

중위수는 전체 관측값을 크기순으로 정렬했을 때, 가운데 위치하는 값이다. 따라서 50%의 관측값이 중위수보다 크거나 같고, 반대로 50%의 관측값이 중위수보다 작거나 같게 된다. 즉, 전체 관측값을 1/2로 나누는 경계값이 중위수이다. 이러한 중위수를 구하는 방법은 다음과 같다.

■ 중위수(median)

관측값을 크기순으로 정렬한 후,

(i) 관측값의 개수(n)가 홀수라면, 중위수는

$$\frac{(n+1)}{2} \text{ 번째 관측값이다.}$$

(ii) 관측값의 개수(n)가 짝수라면, 중위수는

$$\frac{n}{2} \text{ 번째 관측값과 } \frac{n}{2}+1 \text{ 번째 관측값의 평균이다.}$$

보기 2.5

어떤 교과목을 수강한 6명 학생들의 중간고사 성적은 다음과 같다. 학생들의 성적에 대한 표본평균과 중위수를 구하라.

89	78	91	86	76	84

학생들의 성적에 대한 표본평균은 다음과 같다.

$$\bar{x} = \frac{89+78+91+86+76+84}{6} = 84$$

다음으로 중위수를 구하기 위해서는 우선 주어진 자료를 아래와 같이 크기순으로 정렬하여야 한다.

76	78	84	86	89	91

이때, 관측값의 개수는 6개 즉, 짝수이므로 중위수는 가운데 오는 3번째 4번째 관측값의 평균이 된다. 따라서 $(84 + 86)/2 = 85$이다.

보기 2.6

[보기 2.5]의 학생들에 새로운 학생이 추가되었다고 하자. 추가된 학생의 점수가 42점이라면 평균과 중위수는 어떻게 달라지는가?

이제 학생들의 수는 7명이므로 표본평균은 다음과 같아진다.

$$\bar{x} = \frac{42+89+78+91+86+76+84}{7} = 80$$

다음으로 42점이 추가된 자료를 크기순으로 재정렬하면

42	76	78	84	86	89	91

과 같고, 관측값의 개수가 홀수이므로 중위수는 가운데 위치한 4번째 관측값 즉, 84이다.

[보기 2.5]을 통해서 보는 바와 같이 대다수의 관측값들이 평범하고 비슷한 값들이라면 표본평균과 중위수는 비슷한 값을 가지게 된다. 하지만 [보기 2.6]의 경우에서와 같이 아주 크거나 작은 값이 자료 속에 포함되면 평균은 해당 값에 영향을 받아 [보기 2.6]과 같이 때로는 잘못된 중심위치를 나타내기도 한다. 하지만 중위수는 아주 크거나 작은 값에 영향을 적게 받는다. 그러므로 [보기 2.6]의 경우에는 중심위치를 중위수로 표현하는 것이 올바르다.

이처럼 올바른 중심위치를 파악하는 것은 해당 자료의 분포를 파악하는 데 있어서 중요한 요소이다. 그러나 중심위치만으로는 분포를 파악하기에 부족한 점이 많다. 따라서 중심위치와 더불어 해당 중심위치를 기점으로 얼마만큼 흩어져 있는지도 같이 파악하여야만 한다.

표본평균을 중심으로 각각의 관측값들이 얼마나 흩어져 있는지를 파악하기 위해서는 두 값의 차이를 계산하면 된다. 즉, n개의 표본자료를 $x_1, x_2, \cdots, x_n$이라 하고, 이들의 표본평균을 $\bar{x}$라고 하면 $(x_i - \bar{x})$의 값이 각각의 관측값이 표본평균을 중심으로 흩어진 정도를 나타내는 측도가 된다. 이를 편차(deviation)라고 부른다. 그러나 이들 편차의 합은 언제나 0이 되므로, 편차의 제곱합을 구한 후에 관측값의 개수에서 1을 뺀 값으로 나누게 되면 단 하나의

수치로 전체 관측값들이 평균을 중심으로 얼마나 흩어져 있는가를 나타낼 수 있게 된다. 이러한 값을 표본분산이라고 부르고 s^2 으로 표기한다.

■ 표본분산(sample variance)

n 개의 표본자료를 $x_1, x_2, \cdots, x_n$ 이라 하고, 이들의 표본평균을 $\overline{x}$ 라고 하면 표본분산은 다음과 같다.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 .$$

이러한 표본분산의 단위는 언제나 관측값의 측정 단위의 제곱이 되므로, 계산된 수치만으로는 흩어짐의 정도에 대한 크기를 가늠하기가 쉽지 않다. 따라서 표본분산의 양의 제곱근을 통해 관측값의 단위와 일치시키게 되는데, 이를 표본표준편차라고 부르고 s 로 표기한다.

■ 표본표준편차(sample standard deviation)

n 개의 표본자료 $x_1, x_2, \cdots, x_n$ 의 분산을 s^2 이라 하면, 표본표준편차는 다음과 같다.

$$s = +\sqrt{s^2}$$

보기 2.7

[보기 2.5]의 자료에 대해 표본분산과 표본표준편차를 구하라.

명 학생들의 중간고사 평균성적($\overline{x}$)은 84점이므로, 표본분산은 다음의 식에 의해 구할 수 있다.

$$\begin{aligned} s^2 &= \frac{1}{6-1}\sum_{i=1}^{6}(x_i-84)^2 \\ &= \frac{1}{5}\{(89-84)^2+(78-84)^2+\cdots(76-84)^2+(84-84)^2\} \\ &= 35.6 \end{aligned}$$

따라서 표본표준편차 s는 $\sqrt{35.6} \doteq 5.97$이다.

전체 관측값들을 크기순으로 정렬했을 때, 중위수는 전체 관측값을 반 즉, 50%로 나누는 경계값이다. 이러한 개념을 확장하여 전체 관측값을 $(100\times p)$%로 나눌 수 있는 값을 제 $100\times p$ 백분위수라고 한다. 여기서 p는 위치비율을 나타내며, $0\le p\le 1$를 만족한다. 따라서 n개의 관측값이 주어져 있다고 할 때, 제 $100\times p$ 백분위수보다 작거나 같은 관측값의 개수는 np개 이상이 된다. 이러한 제 $100\times p$ 백분위수를 구하는 방법을 정리하면 다음과 같다.

■ 제 $100\times p$ 백분위수(percentile)

관측값을 크기순으로 정렬한 후, 관측값의 개수(n)에 p를 곱했을 때,

(i) np가 정수이면, 제 $100\times p$ 백분위수는

np번째 관측값과 $np+1$번째 관측값의 평균

(ii) np가 정수가 아니면, 제 $100\times p$ 백분위수는

(np의 정수부분에 1을 더한 값)번째 관측값

과 같이 구할 수 있다.

이러한 백분위수들 중에서 제 25백분위수와 제 50백분위수, 제 75백분위수는 사분위수(quartile)들이며, 이들을 각각 제 1사분위수(Q_1)와 제 2사분위수(Q_2), 제 3사분위수(Q_3)라고 부른다. 즉, 사분위수는 전체 자료를 사등분하는 값이다. 추가적으로 제 2사분위수는 중위수이다. 이 때, 제 1사분위수와 제 3사분위수 사이에는 전체 자료의 50%를 포함하게 되며, 이들의 중심에는 중위수가 존재하게 된다. 따라서 중위수를 중심으로 자료의 흩어짐을 파악하기 위한 측도로는 제 1사분위수와 제 3사분위수 사이의 거리를 이용할 수 있는데, 이를 사분위범위라고 하며, 약어로 IQR 이라 표기한다. 이러한 사분위범위는 상위 25%의 관측값과 하위 25%의 관측값을 제외한 중위수 근처의 50% 관측값들에 대한 퍼진 정도를 나타낸다. 따라서 사분위범위는 극단적으로 크거나 작은 값에 영향을 받지 않는 측도이다.

■ 사분위범위(inter-quartile range)

IQR = 제 3사분위수 − 제 1사분위수 = $Q_3 - Q_1$

보기 2.8

[보기 2.5]의 자료에 대해 사분위범위와 제 50백분위수를 구하라

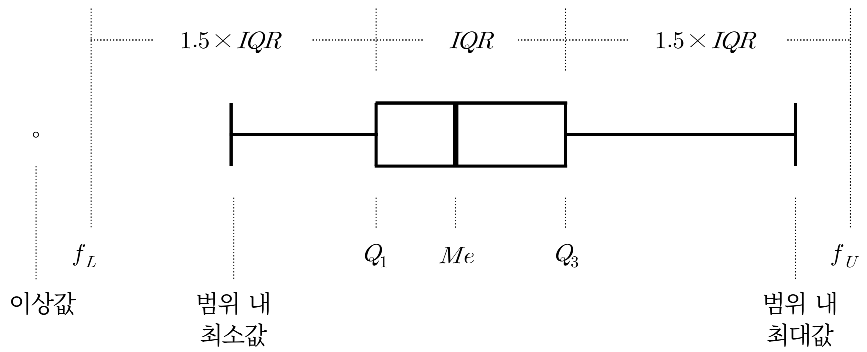
우선 6(= n)개의 관측값을 크기순으로 정렬하면 다음과 같다.

76　　78　　84　　86　　89　　91

사분위수도 백분위수의 한 종류이기에 사분위수도 앞서 설명한 백분위수를 구하는 방법에 의해 구할 수 있다. 즉, 제 1사분위수의 경우에는 $p = 1/4$ 이므로 np 의 값은 1.5이고, 제 3사분위수의 경우에는 $p = 3/4$ 이므로 np 의 값이 4.5이다. 따라서 제 1사분위수와 제 3사분위수는 각각 2번째 관측값과 5번째 관측값인 78과 89가 된다. 마찬가지의 방법으로 제 50백분위수는 $p = 0.5$ 이므로 np 의 값은 3이다. 그러므로 제 50백분위수는 3번째 관측값(84)과 4번째 관측값(86)의 평균인 85가 된다. 이 값은 [보기 2.5]의 결과에서 중위수임을 알 수 있다.

2.5 상자그림

2.3절에서 소개한 히스토그램은 자료가 모여 있는 위치나 자료의 분포에 관한 대략적인 정보를 한눈에 파악할 수 있는 장점이 있지만, 구체적인 수치정보를 쉽게 알아볼 수 없다는 단점이 있다. 따라서 최소값과 제 1사분위수, 중위수, 제 3사분위수, 최대값의 다섯 가지 요약수치 등을 이용하여 그림으로 나타낼 수 있다. 이러한 그림을 상자그림(box plot)이라 부른다. 이러한 상자그림은 [그림 2.4]와 같이 사분위수가 중심이 되어 상자로 표현되며, 상자 좌우로 선을 그려 최소값과 최대값을 나타낸다. 이렇듯 상자 좌우로 선이 그려지기 때문에 때때로 상자그림은 상자-수염그림(box-whisker plot)이라고도 불린다.



[그림 2.4] 상자그림

상자그림을 그리는 방법은 우선 제 1사분위수(Q_1)와 제 3사분위수(Q_3)의 위치에 하나의 네모난 상자를 표현하고, 상자 안 중위수(Me)의 위치에 수직선을 긋는다. 다음으로 사분위범위(IQR)의 1.5배를 계산하고, 상자 양끝으로부터 $1.5 \times IQR$ 크기만큼의 범위를 펼쳐 울타리의 경계값을 계산한다. 즉, 아래쪽 울타리(lower fence, f_L)의 경계값은 $Q_1 - 1.5 \times IQR$이 되고 위쪽 울타리(upper fence, f_U)의 경계값은 $Q_3 + 1.5 \times IQR$이 된다. 이 울타리 범위 내에 포함되는 관측값 중 최소값과 최대값의 위치에 수직선을 긋고 상자와 연결한다. 만약, 울타리 범위를 벗어나는 관측값들이 존재하면 해당 값들을 이상값이라 부르며 상자그림 상에 ◦(혹은 *)로 표시한다.

보기 2.9

어떤 특정 도로를 지나가는 차량의 교통소음을 측정한 값은 아래와 같다. 이 자료에 대한 상자그림을 그려라.

55.9	63.8	57.2	59.8	65.7	62.7	60.8	51.3	61.8	56.0
66.9	56.8	66.2	64.6	59.5	63.1	60.6	62.0	59.4	67.2
63.6	60.5	66.8	61.8	64.8	55.8	55.7	77.1	62.1	61.0
58.9	60.0	66.9	61.7	60.3	51.5	67.0	60.2	56.2	59.4
67.9	64.9	55.7	61.4	62.6	56.4	56.4	69.4	57.6	63.8

상자그림을 그리기에 앞서 주어진 자료를 크기순으로 정렬하면 다음과 같다.

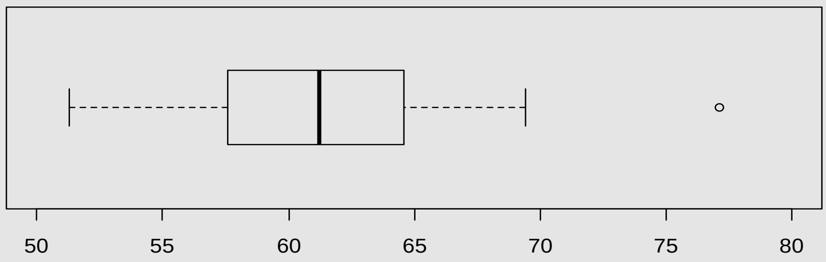
51.3	51.5	55.7	55.7	55.8	55.9	56.0	56.2	56.4	56.4
56.8	57.2	57.6	58.9	59.4	59.4	59.5	59.8	60.0	60.2
60.3	60.5	60.6	60.8	61.0	61.4	61.7	61.8	61.8	62.0
62.1	62.6	62.7	63.1	63.6	63.8	63.8	64.6	64.8	64.9
65.7	66.2	66.8	66.9	66.9	67.0	67.2	67.9	69.4	77.1

관측값의 개수(n)가 50이므로, 제 1사분위수(Q_1)는 $np = 50 \times 0.25 = 12.5$이기에 13번째 관측값인 57.6이고, 제 3사분위수(Q_3)는 $np = 50 \times 0.75 = 37.5$이기에 38번째 관측값인 64.6이다. 이들 값의 위치에 맞게 상자를 표현한다. 다음으로 중위수(Me)는 $np = 50 \times 0.5 = 25$이므로 25번째 관측값(61.0)과 26번째 관측값(61.4)의 평균이 61.2이므로, 해당 위치에 수직선으로 표현한다.

이제 울타리의 경계값을 계산할 차례이다. 사분위범위(IQR)는 $Q_3 - Q_1 = 7$이므로, 아래쪽 울타리(f_L)의 경계값은 $Q_1 - 1.5 \times IQR = 47.1$이고 위쪽 울타리($f_U$)의 경계값은 $Q_3 + 1.5 \times IQR = 75.1$이다. 이들 경계값은 상자그림 상에 표현하지 않는다는 점에 유념하자.

다음으로 상자그림을 완성하기 위해 이상값의 존재 여부를 판단해보자. 자료 속의 최소값은 51.3으로 (47.1, 75.1)의 울타리 안에 포함된다. 최소값을 해당 위치에 수직선으로 표현한다. 그러나 자료의 최대값은 77.1로 울타리 안에 포함되지 않는다. 따라서 77.1은 이상값으로 해당 위치에 ◦로 표현한다.

자료 속에서 2번째로 큰 값은 69.4로 이 값은 울타리 안에 포함된다. 따라서 69.4가 울타리 범위 내의 최대값이 되며, 해당 위치에 수직선을 표현한 후 상자와 연결해주면 상자그림이 완성된다. 이러한 방식으로 그린 상자그림은 다음과 같다.



[그림 2.5] 교통소음 자료에 대한 상자그림

상자그림의 장점은 분포의 다양한 특성이 하나의 그림에 함축되어 있다는 것이다. 다시 말하자면, 상자그림으로부터 자료의 중심위치 및 흩어진 정도 등의 수치정보를 알 수 있는 것뿐만 아니라, 분포의 대칭성과 분포의 집중 정도, 이상값의 존재 여부 등도 알 수 있다.

2.6 R - 프로그램 실습

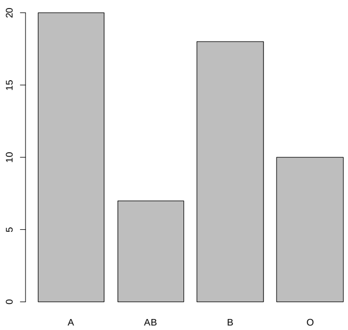
[프로그램 2.1]는 [보기 2.1]의 혈액형 자료에 대한 도수분포표, 막대그래프와 원도표를 그리기 위한 R - 프로그램이다.

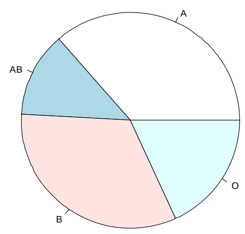
[프로그램 2.1] [보기 2.1]의 혈액형 자료에 대한 도수분포표, 막대그래프와 원도표

```
blood = c("B","A","B","A","A","B","O","A","A","A","O","B","AB","B","AB",
          "AB","A","A","O","AB","O","A","B","O","B","B","A","A","O","A",
          "A","AB","B","B","O","B","B","B","A","AB","A","A","B","O","B",
          "B","O","B","O","B","A","A","AB","A","A")
cnt = table(blood)
prop = prop.table(cnt)
cbind(cnt,prop)
barplot(cnt)
pie(cnt)
```

[결과 2.1] [프로그램 2.1]의 수행 결과

```
   cnt      prop
A   20 0.3636364
AB   7 0.1272727
B   18 0.3272727
O   10 0.1818182
```





[프로그램 2.2]는 [보기 2.3]의 남학생들의 키에 대한 히스토그램을 위한 R - 프로그램이다.

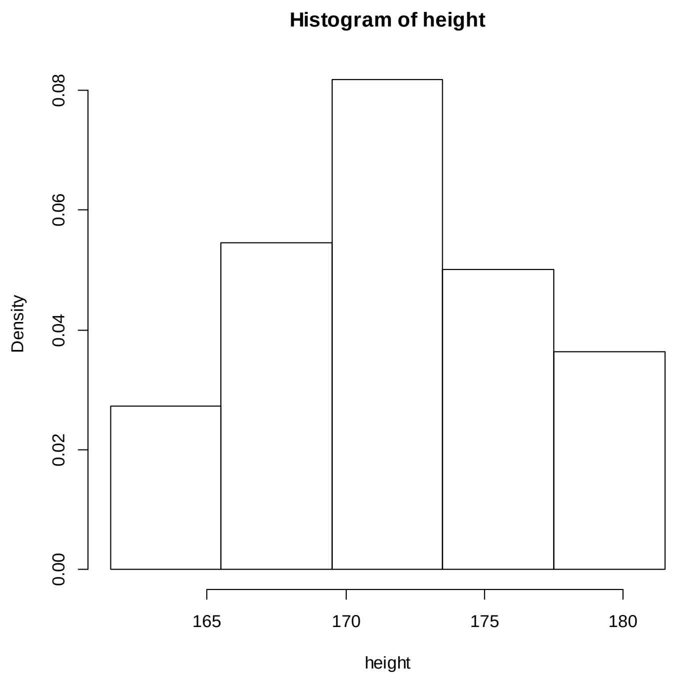
cut 개체에 입력된 수치를 수정하면 [그림 2.3]과 같이 계급의 개수에 따른 히스토그램의 변화를 파악할 수 있다.

[프로그램 2.2] [보기 2.3]의 남학생들 키에 대한 히스토그램

```
height = c(170,178,171,168,173,178,171,174,170,170,175,
        170,169,166,162,170,171,175,175,171,171,170,
        172,179,164,170,181,178,180,177,166,169,168,
        165,163,175,166,178,165,168,167,177,168,177,
        174,174,176,179,169,173,167,170,173,170,162)

cut = c(161.5,165.5,169.5,173.5,177.5,181.5)
hist(height,breaks=cut,probability=T)
```

[결과 2.2] [프로그램 2.2]의 수행 결과



[프로그램 2.3]은 [보기 2.9]의 교통소음 자료에 대해 평균과 분산, 표준편차, 최소값, 최대

값, 제 1사분위수, 중위수, 제 3사분위수를 구하고 상자그림을 그리기 위한 그림이다. 평균을 구하기 위한 R-함수는 mean이고, 분산은 var이며, 표준편차는 sd이다. 그리고 최소값과 최대값, 제 1사분위수, 중위수, 제 3사분위수는 분위수(quantile)이므로 R-함수도 quantile를 이용할 수 있다.

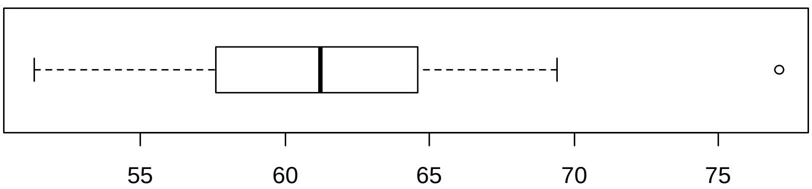
[프로그램 2.3] [보기 2.9]의 교통소음 자료에 대한 기술통계량 및 상자그림

```
noise = c(55.9,63.8,57.2,59.8,65.7,62.7,60.8,51.3,61.8,56.0,
       66.9,56.8,66.2,64.6,59.5,63.1,60.6,62.0,59.4,67.2,
       63.6,60.5,66.8,61.8,64.8,55.8,55.7,77.1,62.1,61.0,
       58.9,60.0,66.9,61.7,60.3,51.5,67.0,60.2,56.2,59.4,
       67.9,64.9,55.7,61.4,62.6,56.4,56.4,69.4,57.6,63.8)

mean(noise)
var(noise)
sd(noise)
quantile(noise,type=2)
boxplot(noise,horizontal=T)
```

[결과 2.3] [프로그램 2.3]의 수행 결과

```
61.374
22.84972
4.780138
  0%  25%  50%  75% 100%
51.3 57.6 61.2 64.6 77.1
```



2.7 연습문제

2.1 통계학을 수강하는 학생들의 생일이 사계절 중 어디에 해당하는지를 조사한 결과이다.

봄	봄	가을	여름	가을	가을	봄	가을	여름
여름	가을	봄	여름	겨울	가을	여름	봄	겨울
여름	가을	가을	가을	여름	겨울	봄	겨울	가을
가을	봄	여름	봄	겨울	여름	겨울		

(1) 도수분포표를 작성하라.

(2) 도수에 대한 막대도표를 그려라.

(3) 원도표를 그리기 위해 계산된 가을에 대한 중심각은 얼마인가?

2.2 다음의 자료에 대하여 각각 표본평균과 중위수를 계산하라.

(1) 수학 시험 성적: 77, 78, 76, 81, 86, 51, 79, 82, 84, 99

(2) A 아파트에 거주하는 가구원수: 1, 5, 2, 3, 2, 1, 4, 1, 3

2.3 연습문제 2.2의 자료에 대하여 각각 표본분산과 표본표준편차를 계산하라.

2.4 연습문제 2.2의 자료에 대하여 각각 사분위범위를 계산하라.

2.5 부산의 35개 영어학원에 등록된 학생 수를 조사한 결과는 다음과 같다.

64	15	21	93	218	43	59
57	28	20	54	52	58	27
100	63	270	94	76	32	175
92	73	183	65	137	177	74
27	28	12	72	281	50	116

(1) 계급의 개수를 5개로 하면 계급구간의 폭은 얼마가 적당한가?

(2) 첫 번째 계급구간의 시작점을 11.5로 하여 도수분포표를 작성하라.

2.6 다음은 아카데미 시상식에서 남우주연상을 받은 배우 30명의 나이를 크기순으로 정렬한 것이다.

18	21	22	25	26	27	29	30	31	33
36	37	41	42	42	47	52	55	57	58
62	64	67	69	71	72	73	74	76	77

(1) 제 40백분위수를 구하라,

(2) 제 78백분위수를 구하라.

2.7 통계와 관련된 세미나에 교사 100명이 참석하였다. 그 중 12명을 대상으로 통계에 대한 인식을 세미나 전과 후에 두 번 조사하였다. 0보다 큰 수는 교사의 통계에 대한 인식이 더 긍정적으로 변한 것을 의미한다. 교사 12명의 통계에 대한 인식의 변화는 다음과 같다.

3, 8, -1, 2, 0, 5, -3, 1, -1, 6, 5, -2

(1) 표본평균을 계산하라.

(2) 표본표준편차를 계산하라.

(3) 중위수를 계산하라.

2.8 다음은 미국 캘리포니아 주에 거주하는 일본계 미국인 27,873명의 연령을 정리한 표이다.

계급구간(세)	도수
0 이상 10 미만	2,568
10 이상 20 미만	2,230
20 이상 30 미만	6,355
30 이상 40 미만	4,181
40 이상 50 미만	3,651
50 이상 60 미만	3,317
60 이상	2,871
합계	27,873

(1) 각 계급의 상대도수를 계산하라.
(2) 히스토그램을 그리기 위하여 각 계급에 해당하는 막대의 높이를 계산하라.
(3) 히스토그램을 그려라.

2.9 다음은 샌프란시스코 지역의 미식축구팀 선수 53명의 체중(단위: 파운드)을 크기순으로 정렬한 것이다.

101	177	178	184	185	185	185	185	188	190
200	205	205	206	210	210	210	212	212	215
215	220	223	228	230	232	241	241	242	245
247	250	250	259	260	260	265	265	270	272
273	275	276	278	280	280	285	285	286	290
290	295	302							

(1) 표본평균을 계산하라.
(2) 중위수를 계산하라.
(3) 제 1사분위수와 제 3사분위수를 계산하라.
(4) 중앙에 위치한 전체의 50%를 차지하는 자료들의 범위를 구하여라.
(5) 상자그림을 그리기 위하여 아래쪽과 위쪽 울타리의 경계값을 계산하라.
(6) 이상값의 존재 여부를 판단하라.
(7) 상자그림을 그려라.

3장

이산확률변수 및 분포

R과 함께하는

통계학의 이해

3.1 사건의 확률

확률(probability)이란 어떠한 실험(experiment)의 결과에 대해 확신하는 정도를 나타낸 수치적 척도(measure)이다. 통계학에서는 실험을 출현 가능한 모든 결과들 중에서 오직 한 가지 결과만이 나타나는 행위를 일컫는다. 이때, 특정 실험에서 나올 수 있는 모든 결과들의 집합을 표본공간(sample space)라고 하며, 일반적으로 S로 나타낸다. 예를 들어, 3개의 공정한 동전을 던지는 행위는 실험이라고 할 수 있으며, 이에 대한 표본공간은 다음과 같다.

$$\mathrm{S} = \{\mathrm{HHH},\ \mathrm{THH},\ \mathrm{HTH},\ \mathrm{HHT},\ \mathrm{TTH},\ \mathrm{THT},\ \mathrm{HTT},\ \mathrm{TTT}\}$$

여기서, H는 각각의 동전이 앞면(head)이 나온 경우를 의미하고 T는 뒷면(tail)이 나온 경우를 의미한다.

우리는 이러한 표본공간의 모든 결과들에 관심을 가지기보다는 특정한 결과들에만 관심을 가지는 경우가 많다. 이처럼 표본공간에 속한 결과들 중에서 어떤 특성을 갖는 결과들의 집합을 사건(event)이라고 하며, 일반적으로 $\mathrm{A}, \mathrm{B}, \cdots$ 로 나타낸다. 더불어 사건 A가 발생할 확률은 $P(\mathrm{A})$로 나타내며, 다음과 같이 정의된다.

$$P(\mathrm{A}) = \frac{\text{사건 A에 속하는 결과의 수}}{\text{표본공간에 속하는 결과의 수}}$$

확률은 그 정의에 의해 언제나 0에서 1사이의 값만을 가지게 된다. 그러므로 어떤 사건의 확률이 0이라 함은 해당 사건이 절대로 일어날 수 없음을 의미하고 확률이 1이라 함은 해당 사건이 항상 일어남을 의미한다. 만약 3개의 공정한 동전을 던져 앞면이 1회 나오는 경우를 사건 A라 정의하면, 사건 A는

$$\mathrm{A} = \{\mathrm{HTT},\ \mathrm{THT},\ \mathrm{TTH}\}$$

가 되며, 따라서 3개의 공정한 동전을 던져 앞면이 1회 나올 확률은

$$P(\mathrm{A}) = \frac{3}{8}$$

과 같이 계산된다.

3.2 확률변수

3.1절에서는 사건을 이용한 확률의 정의를 살펴보았다. 그러나 표본공간의 크기가 커질수록 표본공간을 정의하기는 쉽지가 않다. 이럴 때 이용할 수 있는 것이 확률변수(random variable)이다. 확률변수는 표본공간에 속하는 각각의 결과들에 대해 실수값을 대응시켜 주는 함수를 의미한다. 예를 들어 확률변수 X를 3개의 공정한 동전을 던졌을 때 앞면이 나온 횟수라고 정의한다면, 확률변수 X는 0(회), 1(회), 2(회), 3(회)의 4개의 값을 가질 수 있으며 이에 대응되는 사건들은 다음과 같다.

확률변수 X의 값	X의 값에 대응되는 사건
0	{TTT}
1	{HTT, THT, TTH}
2	{HHT, HTH, THH}
3	{HHH}

즉, 확률변수는 각각의 실험 결과들을 수치적으로 표현하는 방법이라 할 수 있다. 이때, 일반적으로 확률변수는 $X, Y, \cdots$와 같이 대문자로 나타내고, 확률변수가 가지는 특정값은 $x, y, \cdots$와 같이 소문자로 나타낸다.

확률변수는 가질 수 있는 값에 따라 이산확률변수(discrete random variable)와 연속확률변수(continuous random variable)로 구분한다. 이산확률변수는 확률변수가 가질 수 있는 값들을 일일이 지칭할 수 있는 확률변수를 의미한다. 따라서 이산확률변수는 확률변수가 가질 수 있는 값들의 수가 설령 무한 개라 하더라도 셀 수 있다. 이와는 달리, 연속확률변수는 확률변수가 특정 구간의 모든 값을 다 가질 수 있기 때문에 가질 수 있는 값들을 일일이 지칭할 수 없는 확률변수를 의미한다. 이 장에서는 이산확률변수의 특징 및 확률분포를 다루고 4장에서 연속확률변수를 다루고자 한다.

3.3 이산확률변수의 확률분포함수

3개의 공정한 동전을 던졌을 때 앞면이 나온 횟수를 확률변수 X라 하면, 확률변수 X는 앞 절에서 확인한 바와 같이 4개의 값을 가질 수 있으므로 이산확률변수이며 각각의 값을 가질 때의 확률은 다음과 같다.

x	0	1	2	3	합계
$P(X=x)$	1/8	3/8	3/8	1/8	1

이처럼 확률변수가 가지는 값과 그 값을 가질 확률을 정해주는 규칙 또는 관계를 확률분포(probability distribution)라고 부른다. 이러한 확률분포는 언제나 나름대로 일련의 규칙을 가지게 되므로, 함수를 이용하여 표현 가능하다. 이를 확률분포함수(probability distribution function)라고 한다.

일반적으로 이산확률변수는 가질 수 있는 값들을 일일이 지칭할 수 있기 때문에 이산확률변수의 확률분포함수 $f(x)$는

$$f(x)=P(X=x)$$

와 같이, 확률변수 X가 특정한 값 x를 갖게 되는 확률을 나타낸다. 따라서 이산확률변수의 확률분포함수는 다음의 두 가지 성질을 만족한다.

■ 이산확률분포함수의 성질

- 모든 x 값에 대해 $0 \le f(x) \le 1$
- $\sum_{x} f(x)=1$

3.4 확률변수의 기대값과 표준편차

2장에서 우리는 측정된 관측값들의 중심위치를 파악하기 위해 표본평균을 공부하였다. 이와 유사한 개념으로 확률변수도 가질 수 있는 값이 많으므로 그 값들의 중심위치를 파악할 필요가 있다. 이 중 확률변수가 가질 수 있는 값들에 대한 확률분포 상의 중심위치를 그 확률변수의 기대값(expected value)이라고 한다. 일반적으로 확률변수 X의 기대값은 $E(X)$로 나타내며, 이산확률변수의 기대값은 다음과 같이 계산할 수 있다.

$$E(X) = \sum_x x f(x) \tag{3.1}$$

보기 3.1

어떤 축구팀은 한 시즌 총 38주 중에 매주 1회 혹은 2회의 경기 일정이 잡혀 있다고 한다. 1주일에 1회의 경기가 있는 비율은 0.5, 2회의 경기가 있는 비율은 0.3이며, 나머지는 경기가 없는 휴식주간이라고 한다. 이 팀은 평균적으로 매주 몇 경기를 수행하여야 하는가?

이 팀이 매주 뛰어야 하는 경기 수를 확률변수 X라고 하면, 확률변수 X의 확률분포는 다음과 같다.

x	0	1	2	합계
$f(x)$	0.2	0.5	0.3	1

확률변수 X는 3개의 값만을 가질 수 있으므로 이산확률변수이며, 따라서 이 팀이 평균적으로 매주 수행해야 하는 경기 수 즉, 경기 수에 대한 기대값은 다음과 같이 계산할 수 있다.

$$E(X) = \sum_x x f(x) = 0 \times 0.2 + 1 \times 0.5 + 2 \times 0.3 = 1.1$$

그러므로 이 팀은 평균적으로 매주 1.1회의 경기를 수행하여야 한다.

이처럼 기대값은 종종 평균을 의미하게 된다. 그러나 기대값이 의미하는 평균은 2장에서 배운 표본평균과 다르다. 표본평균은 단기적 실험 결과에 대한 평균을 의미하는 반면, 기대값은 장기적 실험 결과에 대한 평균을 의미한다. 예를 들어 2개의 공정한 동전을 던지는 실험을 생각해 보자. 이 실험을 실제로 10회 반복한 경우, 각각의 실험에서 앞면이 나온 횟수를 기록하면

실험 번호	1	2	3	4	5	6	7	8	9	10
앞면이 나온 횟수	0	1	1	2	0	1	1	0	1	2

와 같을 수 있다. 따라서 총 10회의 실험에서 2개의 공정한 동전에 대해 앞면이 나온 평균 횟수(표본평균)는

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{0+1+1+2+0+1+1+0+1+2}{10} = 0.9$$

와 같다. 이러한 표본평균은 10회의 실험을 할 때마다 다르게 계산될 수 있다. 하지만 기대값은 이처럼 단 10회의 실험에 대한 평균을 의미하지 않는다. 기대값은 동일한 실험을 무수히 반복했을 때의 평균을 의미한다. 공정한 동전 1개를 10회 던졌을 때 앞면이 8번 나왔다고 해서 동전을 던지면 앞면이 나올 확률이 0.8이라고 하지 않는다. 일반적으로 공정한 동전 1개에 대해 앞면이 나올 확률을 0.5라고 하는 것은 해당 동전을 무수히 많이 던졌을 때 앞면이 나오는 비율이 0.5라는 것을 의미한다. 이처럼 기대값도 동일한 실험을 무수히 반복했을 때의 평균 즉, 확률분포 상의 중심위치를 의미한다. 따라서 공정한 동전 2개를 던졌을 때 앞면이 나온 횟수를 확률변수 X라 하면, 확률변수 X에 대한 확률분포는

x	0	1	2
$f(x)$	1/4	2/4	1/4

와 같으므로 공정한 동전 2개를 던진 경우 앞면이 나온 횟수에 대한 기대값은

$$E(X) = \sum_x x f(x) = 0 \times \frac{1}{4} + 1 \times \frac{2}{4} + 2 \times \frac{1}{4} = 1$$

와 같이 계산된다. 이는 공정한 동전 2개를 던지면 평균적으로 1개의 동전은 앞면이 나온다는 것을 의미한다. 이러한 기대값은 어떠한 실험을 행할 경우 나올 수 있는 모든 결과들에 대한 평균이므로 이는 모집단 전체에 대한 평균 즉, 모평균(population mean)과 같다. 일반적으로 통계학에서는 모평균을 μ로 나타낸다. 이러한 기대값은 다음과 같은 성질을 만족한다.

■ 기대값의 성질

a와 b는 상수이고 X는 임의의 확률변수라 할 때,

(i) $E(X) = \mu$
(ii) $E(a) = a$
(iii) $E(aX) = aE(X) = a\mu$
(iv) $E(aX \pm b) = E(aX) \pm E(b) = a\mu \pm b$

보기 3.2

공정한 주사위 1개를 던져 윗면에 나타난 수에 100원을 곱한 금액을 돌려주는 게임을 한다고 가정하자. 이 게임을 하는 경우 한 게임당 평균적으로 얼마의 이득을 취할 수 있는가? 단, 게임 한 판당 200원의 비용을 지불해야 한다.

공정한 주사위 1개를 던졌을 때 윗면에 나타난 수를 확률변수 X라고 하면, 확률변수 X의 확률분포는 다음과 같다.

x	1	2	3	4	5	6	합계
$f(x)$	1/6	1/6	1/6	1/6	1/6	1/6	1

따라서 확률변수 X의 기대값은

$$E(X)=\sum_x xf(x)=1\times\frac{1}{6}+2\times\frac{1}{6}+3\times\frac{1}{6}+4\times\frac{1}{6}+5\times\frac{1}{6}+6\times\frac{1}{6}=3.5$$

와 같다. 그러나 우리의 실제 관심 사항은 한 게임당 취할 수 있는 이득이므로, 이를 확률변수 Y라 둔다면 확률변수 Y와 확률변수 X의 관계는

$$Y=100\times X-200$$

의 관계를 만족하며 확률분포는 다음과 같다.

y	-100	0	100	200	300	400	합계
$f(x)$	1/6	1/6	1/6	1/6	1/6	1/6	1

그러므로 한 게임당 평균적으로 취할 수 있는 이득 즉, 확률변수 Y의 기대값은

$$E(Y)=(-100)\times\frac{1}{6}+0\times\frac{1}{6}+100\times\frac{1}{6}+200\times\frac{1}{6}+300\times\frac{1}{6}+400\times\frac{1}{6}=150$$

과 같이 계산 가능하다. 이는 기대값의 성질을 이용하여

$$E(Y)=E(100\times X-200)=E(100\times X)-E(200)=100\times 3.5-200=150$$

과 같은 방법으로도 계산할 수 있다.

확률변수의 기대값은 중심위치 즉, 모평균을 나타내므로, 이러한 기대값을 중심으로 확률변수가 가질 수 있는 값의 흩어짐을 파악하기 위해서는 분산과 표준편차를 이용할 수 있다. 확률변수 X의 분산은 $Var(X)$로 나타내며, 이산확률변수의 분산은

$$Var(X)=\sum_x (x-\mu)^2 f(x) \tag{3.2}$$

와 같이 계산할 수 있다. 물론 확률변수 X의 표준편차는 분산의 양의 제곱근이므로 다음과 같이 계산할 수 있다.

$$\sqrt{Var(X)} = \sqrt{\sum_x (x-\mu)^2 f(x)}$$

확률변수의 분산을 좀 더 수리적으로 표현하자면, 식 (3.1)과 비교하였을 때 x 대신에 $(x-\mu)^2$ 이 사용되었음을 알 수 있다. 여기서 $E(X) = \mu$ 이므로 식 (3.2)의 표현은

$$Var(X) = E[\,(X-E(X))^2\,] \tag{3.3}$$

와 같다. 즉, 확률변수 X가 가질 수 있는 값에서 기대값을 뺀 후 제곱한 값들에 대해 다시 기대값을 계산한 것이다. 이 때, 기대값 $E(X) = \mu$ 는 상수이므로 식 (3.3)은 기대값의 성질에 의해 다음과 같이 표현할 수 있다.

$$\begin{aligned} Var(X) &= E[\,(X-E(X))^2\,] \\ &= E[\,X^2 - 2\mu X + \mu^2\,] \\ &= E(X^2) - 2\mu E(X) + E(\mu^2) \\ &= E(X^2) - \mu^2 \\ &= E(X^2) - E(X)^2 \end{aligned} \tag{3.4}$$

따라서 때로는 식 (3.4)의 계산방식이 분산의 계산에 편의성을 제공할 수 있다.

보기 3.3

[보기 3.1]의 축구팀이 1주일에 수행하여야 하는 경기 수의 분산과 표준편차는 얼마인가?

이 팀이 매주 뛰어야하는 경기 수를 확률변수 X라고 하면, 확률변수 X의 확률분포는

x	0	1	2	합계

$f(x)$	0.2	0.5	0.3	1

와 같고, 확률변수 X의 기대값은 1.1임을 이미 확인한 바 있다.
식 (3.2)의 방식에 따라 확률변수의 분산을 계산하기 위해서는 다음과 같은 계산과정이 필요하다.

x	0	1	2
$f(x)$	0.2	0.5	0.3
$(x-\mu)^2 f(x)$	$(0-1.1)^2\times 0.2$	$(1-1.1)^2\times 0.5$	$(2-1.1)^2\times 0.3$

따라서 확률변수 X의 분산은 다음과 같이 계산할 수 있다.

$$\begin{aligned} Var(X) &= \sum_x (x-\mu)^2 f(x) \\ &= (0-1.1)^2\times 0.2 + (1-1.1)^2\times 0.5 + (2-1.1)^2\times 0.3 \\ &= 0.49 \end{aligned}$$

다음으로 식 (3.4)의 방식에 따라 확률변수의 분산을 계산해 보자. 이를 위해서는 $E(X^2)$ 즉, 확률변수 X^2에 대한 기대값을 계산하여야 한다. 확률변수 X^2은 이 팀이 매주 뛰어야 하는 경기 수의 제곱이므로, 확률변수가 가질 수 있는 값과 확률분포는 다음과 같다.

x^2	0	1	4	합계
$f(x^2)$	0.2	0.5	0.3	1

여기서 확률변수 X^2의 확률분포 $f(x^2)$은 확률변수 X의 확률분포 $f(x)$와 같음을 알 수 있다. 따라서 확률변수 X^2의 기대값은

$$E(X^2) = \sum_{x^2} x^2 f(x^2) = \sum_x x^2 f(x) = 0^2\times 0.2 + 1^2\times 0.5 + 2^2\times 0.3 = 1.7$$

과 같이 계산할 수 있다. 그러므로 확률변수 X의 분산은 다음과 같이 계산할 수 있다.

$$Var(X) = E(X^2) - E(X)^2 = 1.7 - 1.1^2 = 0.49$$

즉, 식 (3.2)와 식 (3.4) 모두 같은 결과임을 알 수 있다. 더불어 확률변수 X의 표준편차는 분산의 양의 제곱근이므로 다음과 같다.

$$\sqrt{Var(X)} = \sqrt{0.49} = 0.7$$

일반적으로 확률변수의 기대값이 모평균을 나타내듯이 분산은 모집단의 분산 즉, 모분산을 나타내며 기호로는 σ^2 으로 표현한다. 이러한 분산은 다음과 같은 성질을 만족한다.

■ 분산의 성질

a와 b는 상수이고 X는 임의의 확률변수라 할 때,

(i) $Var(X) = \sigma^2$

(ii) $Var(a) = 0$

(iii) $Var(aX) = a^2 Var(X) = a^2\sigma^2$

(iv) $Var(aX \pm b) = Var(aX) = a^2\sigma^2$

보기 3.4

[보기 3.2]의 게임을 하는 경우 한 게임당 취할 수 있는 이득의 분산과 표준편차는 각각 얼마인가?

공정한 주사위 1개를 던졌을 때 윗면에 나타난 수를 확률변수 X라고 하고 한 게임당 취할 수 있는 이득을 확률변수 Y라고 하면, 확률변수 X의 Y의 확률분포는

x	1	2	3	4	5	6	합계
y	-100	0	100	200	300	400	
$f(x)=f(y)$	1/6	1/6	1/6	1/6	1/6	1/6	1

와 같고, 확률변수 X와 Y의 기대값은 각각 3.5와 150임을 확인한 바 있다. 따라서 한 게임당 취할 수 있는 이득의 분산 즉, 확률변수 Y의 분산은

$$Var(Y) = E(Y^2) - E(Y)^2$$

와 같이 식 (3.4)를 이용하여 직접적으로 계산할 수도 있지만, 확률변수 X와 Y의 관계 및 분산의 성질을 이용하여

$$Var(Y) = Var(100 \times X - 200) = Var(100 \times X) = 100^2 \times Var(X)$$

와 같은 방법으로도 계산 가능하다. 여기서,

$$E(X^2) = 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + 3^2 \times \frac{1}{6} + 4^2 \times \frac{1}{6} + 5^2 \times \frac{1}{6} + 6^2 \times \frac{1}{6} = \frac{91}{6}$$

이므로 확률변수 X의 분산은 다음과 같다.

$$Var(X) = E(X^2) - E(X)^2 = \frac{91}{6} - 3.5^2 \doteq 2.9167$$

따라서 확률변수 Y의 분산은

$$Var(Y) = 100^2 \times Var(X) \doteq 100^2 \times 2.9167 = 29167$$

와 같으며, 확률변수 Y의 표준편차는 다음과 같다.

$$\sqrt{Var(Y)} \doteq 170.78$$

3.5 이항분포

3.3절에서 확률분포는 언제나 나름대로 일련의 규칙을 가지게 된다는 것을 언급한 바 있다. 이미 많은 확률분포들은 그 규칙들이 밝혀져 이름이 부여된 것들이 많다. 이산형 확률변수가 가지는 확률분포 중에서 일상생활에서 가장 활용도가 높은 확률분포는 이항분포(binomial distribution)이다. 따라서 이번 절에서는 이항분포의 특징과 그 활용방안을 소개하고자 한다. 우선, 이항분포를 배우기에 앞서 우선적으로 베르누이 시행(Bernoulli trial)에 대해 알아보자. 베르누이 시행이란 단 1회의 실험을 지칭하며, 그 특징은 다음과 같다.

■ 베르누이 시행의 특징

- 각 시행은 성공(success, S)과 실패(fail, F)의 두 가지 결과만을 갖는다.
- 각 시행에서 성공할 확률은 $P(S)=p$, 실패할 확률은 $P(F)=1-p$로 매 시행마다 동일하다.
- 각 시행은 상호 독립으로 각각의 시행이 다른 시행의 결과에 영향을 미치지 않는다.

이러한 베르누이 시행을 n번 반복한 경우 성공 횟수를 확률변수 X라 하면, 이 확률변수 X의 확률분포는 이항분포를 따르게 된다. 성공확률이 p인 베르누이 시행을 n번 시행한 경우 성공 횟수를 나타내는 확률변수 X의 확률분포함수는

$$f(x)=P(X=x)={}_n\mathrm{C}_x\, p^x(1-p)^{n-x}, \ x=0,1,\cdots,n$$

와 같다. 예를 들어, 공정한 동전 3개를 던졌을 때 앞면이 나온 횟수를 확률변수 X라 하면, 확률변수 X에 대한 확률분포는

x	0	1	2	3	합계
$P(X=x)$	1/8	3/8	3/8	1/8	1

와 같음을 3.2절에서 이미 확인한 바 있다. 여기서 공정한 동전 1개를 던지는 행위는 성공(앞

면, H)과 실패(뒷면, T)로 나눌 수 있으며 성공확률 즉, 앞면이 나올 확률이 0.5로 정해져 있다. 더불어 동전 1개를 반복해서 던지더라도 매 시행의 성공확률은 독립적이므로 동전 1개를 던지는 행위는 베르누이 시행이라고 할 수 있다. 따라서 공정한 동전 3개를 던지는 실험은 성공확률(p)이 0.5인 베르누이 시행을 $n=3$ 번 반복한 경우이므로, 앞면이 나온 횟수(성공횟수)를 확률변수 X라 하면, 확률변수 X는 이항분포를 따르게 된다. 확률분포함수를 앞면이 1번 나올 확률로 계산해 보자. 우선 공정한 동전 3개를 던져 앞면이 1번 나오려면 뒷면이 2번 나오는 사건이 동시에 발생하여야 한다. 따라서 앞면이 나올 확률 p와 뒷면이 나올 확률 $(1-p)$를 이용하여 기호화하면

$$p\times(1-p)\times(1-p)=p^1(1-p)^{3-1}$$

와 같이 표현 가능하다. 그런데 이러한 사건은 3개의 동전 중에서 1개의 동전만 앞면이 나오는 경우의 수 즉, ${}_3\mathrm{C}_1=3$ 가지의 경우가 있다. 따라서 공정한 동전 3개를 던져 앞면이 1번 나올 확률은 이항분포의 확률분포함수를 이용하여

$$f(1)=P(X=1)={}_3\mathrm{C}_1\times\left(\frac{1}{2}\right)^1\times\left(1-\frac{1}{2}\right)^{3-1}=\frac{3}{8}$$

과 같이 계산할 수 있다. 이처럼 임의의 이산형 확률변수가 이항분포를 따른다는 것을 파악할 수만 있다면, 그 확률변수가 가지는 각각의 값에 대한 확률을 간편하게 계산할 수 있다.

일반적으로 임의의 이산형 확률변수 X가 n 번 시행에서 성공확률이 p 인 이항분포를 따른다는 표현은

$$X\sim \mathrm{Bin}\,(n,p)$$

와 같이 나타낸다. 즉, 공정한 동전 3개를 던졌을 때 앞면이 나온 횟수를 나타내는 확률변수 X는 시행 횟수가 $n=3$ 이고 성공확률이 $p=0.5$ 이므로

$$X \sim \mathrm{Bin}(3, 0.5)$$

와 같이 표현한다.

만약 주어진 상황에서 정의한 확률변수가 이항분포를 따른다는 것을 파악할 수만 있다면, 그 확률변수에 대한 기대값과 분산도 손쉽게 계산할 수 있다. n 번의 시행에서 성공확률이 p 인 이항분포를 따르는 이산형 확률변수 X의 기대값과 분산은 다음과 같다.

■ 이항분포의 기대값과 분산

성공 횟수를 나타내는 확률변수 X가 $X \sim \mathrm{Bin}(n, p)$ 일 때,

(i) $E(X) = np$

(ii) $Var(X) = np(1-p)$

예를 들어, 공정한 동전 100개를 던졌을 때 앞면이 나온 횟수를 확률변수 X라고 하면, 확률변수 X는 시행 횟수가 $n = 100$ 이고 성공확률이 $p = 0.5$ 인 이항분포를 따르므로

$$X \sim \mathrm{Bin}(100, 0.5)$$

와 같이 표현 가능하며, 이 확률변수 X의 기대값과 분산, 표준편차는 각각

$$E(X) = np = 100 \times 0.5 = 50,$$

$$Var(X) = np(1-p) = 100 \times 0.5 \times 0.5 = 25,$$

$$\sqrt{Var(X)} = \sqrt{np(1-p)} = \sqrt{25} = 5$$

와 같이 계산된다. 따라서 실제로 공정한 동전 100개를 던졌을 때 앞면이 나온 횟수를 파악해 보면 50± 5회 즉, 45회부터 55회 사이가 흔히 나타나게 됨을 알 수 있다.

보기 3.5

대도시의 시민 중 30%만이 대중교통에 만족한다고 한다. 만약 이 도시의 시민 20명을 임의로 선택했을 때, 이들 중 3명 미만의 시민이 대중교통에 만족할 확률과 6명의 시민이 만족할 확률을 구하라.

대중교통에 만족하면 성공, 만족하지 않으면 실패라고 한다면 각각의 시민에 대한 성공확률 즉, 대중교통에 만족할 확률은 0.3이다. 그리고 임의로 20명을 선택했으므로 이들 독립적 선택을 한다고 가정할 수 있을 것이다. 따라서 선택된 20명 중 대중교통에 만족하는 시민의 수를 확률변수 X라고 하면, 이러한 확률변수 X는 다음과 같은 이항분포를 따르게 된다.

$$X \sim \mathrm{Bin}(20, 0.3)$$

1) 이들 중 3명 미만의 시민이 대중교통에 만족할 확률은

$$\begin{aligned} P(X<3) &= P(X=0)+P(X=1)+P(X=2) \\ &= {}_{20}\mathrm{C}_0 \times 0.3^0 \times 0.7^{20-0} + {}_{20}\mathrm{C}_1 \times 0.3^1 \times 0.7^{20-1} + {}_{20}\mathrm{C}_2 \times 0.3^2 \times 0.7^{20-} \\ &\doteq 0.0355 \end{aligned}$$

2) 6명의 시민이 대중교통에 만족할 확률은

$$P(X=6) = {}_{20}\mathrm{C}_0 \times 0.3^6 \times 0.7^{20-6} = 0.1916$$

보기 3.6

어떤 바이러스 질환은 감염자와 접촉을 하게 되면 감염되며, 건강한 사람이 감염자와 한 번 접촉하였을 때 감염될 확률은 20%라고 한다. 감염자가 임의의 건강한 사람 5명과 접촉했을 때, 5명 모두 감염될 확률은 얼마인가? 그리고 감염자 수의 기대값과 분산, 표준편차는 얼마인가?

감염자와 접촉한 5명의 건강한 사람 중 감염된 사람의 수를 확률변수 X라고 하면, 확률변수 X는 다음과 같은 이항분포를 따른다.

$$X \sim \mathrm{Bin}(5, 0.2)$$

1) 이들 모두가 감염되었을 확률은

$$P(X=5) = {}_5C_5 \times 0.2^5 \times 0.8^{5-5} \doteq 0.0003$$

2) 감염자 수의 기대값과 분산, 표준편차는?

$$E(X) = np = 5 \times 0.2 = 1$$

$$Var(X) = np(1-p) = 5 \times 0.2 \times 0.8 = 0.8$$

$$\sqrt{Var(X)} = \sqrt{np(1-p)} = \sqrt{0.8} \doteq 0.8944$$

3.6 R-프로그램 실습

[프로그램 3.1]은 [보기 3.5]의 대중교통에 만족하는 시민의 수에 대해 이항분포를 활용한 R-프로그램을 보여주고 있다. R-프로그램의 내장함수 중 dbinom과 pbinom을 이용하면 이항분포에서의 확률을 쉽게 계산할 수 있다. 구문형식은

```
dbinom(x, size, prob)
pbinom(x, size, prob, lower.tail = TRUE)
```

이며, 여기서 x는 확률변수 X의 특정값을 의미하고, size는 시행횟수(n)를 의미하며, prob는 성공확률(p)을 의미한다. 그리고 pbinom 함수의 lower.tail은 확률계산의 방향을 선택하는 옵션으로 그 값을 TRUE(또는 T)로 설정하면 $P(X \leq x)$가 계산되고, FALSE(또는 F)로 하면 $P(X \geq x)$가 계산된다.

[프로그램 3.1] [보기 3.5] 대중교통에 만족하는 시민의 수에 대한 확률

```
pbinom(2,size=20,prob=0.3,lower.tail=T)
dbinom(6,size=20,prob=0.3)
```

[결과 3.1] [프로그램 3.1]의 수행 결과

```
0.03548313
0.191639
```

만약, 10명 이상의 시민이 대중교통에 만족할 확률을 구하고자 한다면 [프로그램 3.2]와 같이 작성하면 된다.

[프로그램 3.2] [보기 3.5] 대중교통에 만족하는 시민의 수에 대한 확률

```
pbinom(10,20,0.3,lower.tail=F)
```

[결과 3.2] [프로그램 3.2]의 수행 결과

```
0.01714482
```

3.7 연습문제

3.1 한 벤처투자가가 100만 원을 투자하기 위하여 선택할 수 있는 3가지 회사가 있다. A회사는 500만 원의 수익을 낼 확률이 10%, 100만 원의 수익을 낼 확률이 30%, 100만 원을 손해 볼 확률이 60%이다. B회사는 300만 원의 수익을 낼 확률이 20%, 100만 원의 수익을 낼 확률이 40%, 100만 원을 손해 볼 확률이 40%이다. C회사는 600만 원의 수익을 낼 확률이 10%, 아무런 이익도 손해도 보지 않을 확률이 70%, 100만 원을 손해 볼 확률이 20%이다. 여기서 각 회사에 투자하여 얻을 수 있는 이익을 X라 하자.

(1) 각 회사에 대한 X의 확률분포를 구하라.

(2) 각 회사에 대한 X의 기대값을 구하라. 가장 높은 기대 수익을 갖는 투자는 어느 회사인가?

(3) 가장 안전한 투자와 가장 위험한 투자는 어느 회사인가?

3.2 부산에 거주하는 기혼 성인 2만 명을 대상으로 자녀수를 조사한 결과는 다음과 같다. 여기서 자녀수를 X라 하자.

x	0	1	2	3	4	5	6
$P(X=x)$	0.10	0.20	0.30		0.10	0.05	0.05

(1) 자녀수가 3명일 확률을 구하라.

(2) X의 기대값과 표준편차를 구하라.

(3) 자녀수가 2 - 3명일 가능성이 큰가? 4 - 6명일 가능성이 큰가? 그 이유는?

3.3 색상이 다른 주사위 6개를 던져서 1이 나오는 주사위의 개수에 관심이 있다.

(1) 확률변수 X를 정의하라.

(2) X가 취할 수 있는 값은?

(3) X는 어떤 분포를 가지는가?

(4) X의 기대값과 표준편차를 구하라.

(5) 6개의 주사위가 모두 1이 나올 확률을 구하라.

3.4 통계학 학사 학위를 얻기 위해 걸리는 기간은 다음과 같다.

x	3	4	5	6	7
$P(X=x)$	0.05	0.40	0.30	0.15	0.10

(1) 확률변수 X를 정의하라.

(2) $X=0,1,2$에 대하여 확률분포함수 $P(X)$가 존재하지 않는 것은 무엇을 의미하는가?

(3) 평균적으로 통계학 학사 학위를 얻기 위해 얼마나 걸리는가?

3.5 롯데 자이언츠 야구팀이 13년 동안 1,034경기 중 382경기를 이겨 승률이 0.3694로 주어져 있다. 다음 달에 12경기를 가질 예정이다. 여기서 X는 다음 달에 이긴 경기 수이다.

(1) 다음 달에 몇 경기를 이길 것으로 기대되는가?

(2) 6경기를 이길 확률을 구하라.

(3) 최소 5경기를 이길 확률을 구하라.

3.6 펜싱 선수의 60%는 주무기가 포일(foil)이다. 펜싱 선수 25명을 무작위로 조사하였다. 여기서 주무기가 포일이 아닌 선수의 수에 관심이 있다.

(1) 확률변수 X를 정의하라.

(2) X는 어떤 분포를 가지는가?

(3) 주무기가 포일이 아닌 선수는 몇 명으로 기대되는가?

(4) 25명 모두 주무기가 포일이 아닐 수 있는지 수치적으로 설명하라.

3.7 공정한 주사위를 사용하여 다음의 게임을 한다고 가정하자. 1부터 6까지 숫자 중 하나를 선택하여 1만 원을 걸고, 주사위를 세 번 던진다. 만일 선택한 숫자가 한 번도 나오지 않으면 처음 낸 1만 원을 잃는다. 만일 선택한 숫자가 한 번 나오면 처음 낸 1만 원을 되돌려 받고 1만 원을 추가로 더 받으며, 선택한 숫자가 두 번 나오면 처음 낸 1만 원을 되돌려 받고 2만 원을 추가로 더 받는다. 마찬가지로 선택한 숫자가 세 번 나오면 처음 낸 1만 원을 되돌려 받고 3만 원을 추가로 더 받는다. 여기서 X는 선택한 숫자가 나오는 횟수이고, Y는 게임에서 얻는 금액이라 하자.

(1) X는 어떤 분포를 가지는가?

(2) Y가 취할 수 있는 값은?

(3) X와 Y에 대한 확률분포표를 작성하라.

(4) X와 Y에 대한 기대값을 구하라.

(5) 다음 두 가지 방법으로 X의 분산을 구하여 비교하라.

i) 확률분포표를 이용　ii) 이항분포의 분산 공식 이용

3.8 서울의 한 비디오 대여점에 방문하는 사람들은 한 번에 한 개 이상의 DVD를 빌려 간다. 한 손님당 빌려가는 DVD의 개수에 대한 확률은 다음과 같다.

x	0	1	2	3	4	5
$P(X=x)$	0.03	0.50	0.24		0.07	0.04

(1) 한 손님이 DVD를 3개 빌려 갈 확률은?

(2) 한 손님이 DVD를 적어도 4개 빌려 갈 확률은?

(3) 한 손님이 DVD를 최대 2개 빌려 갈 확률은?

(4) 다음은 부산의 다른 비디오 대여점에 방문하는 사람들이 빌려가는 DVD의 개수에 대한 확률이다. 두 지역의 비디오 대여점 중 한 사람당 빌려가는 DVD의 개수가 더 많은 곳은 어느 곳인가?

y	0	1	2	3	4	5
$P(Y=y)$	0.35	0.25	0.20	0.10	0.05	0.05

3.9 한 벌의 카드에서 카드 1장을 선택하고, 동전을 던지는 게임을 한다. 한 벌의 카드는 52장이고, 그 중 12장의 카드는 '페이스카드(King, Queen, Jack)'이다. 동전은 앞면과 뒷면이 나올 확률이 동일하다. 만일 페이스카드를 선택하고 동전 앞면이 나오면 6만 원을 얻고, 페이스카드를 선택하고 동전 뒷면이 나오면 2만 원을 얻는다. 그리고 페이스카드를 선택하지 않으면 동전의 앞뒤와 관계없이 2만 원을 잃는다. 이 게임에서 손익에 대한 기대값을 구하고, 돈을 얻기 위해 이 게임을 하는 것이 옳은지를 말하라.

3.10 한 대학신문 기자는 올해 학교 축제에 참여할 것인가를 알아보기 위하여 12명의 학생을 무작위로 조사하기로 하였다. 지난해에 따르면, 학생의 18%가 학교 축제에 참여한 것으로 알려져 있다. 여기서 학교 축제에 참여할 학생의 수를 X라 하자.

(1) 12명 중 몇 명의 학생이 축제에 참여할 것으로 기대되는가?

(2) 최대 4명의 학생이 축제에 참여할 확률을 구하라.

(3) 적어도 3명 이상의 학생이 축제에 참여할 확률을 구하라.

4장

연속확률변수 및 분포

R과 함께하는

통계학의 이해

4.1 연속확률변수의 확률분포함수

연속확률변수는 확률변수가 특정 구간의 모든 값을 다 가질 수 있기 때문에 가질 수 있는 값들을 일일이 지칭할 수 없는 확률변수를 의미함을 3.2절에서 언급한 바 있다. 따라서 연속확률변수 X의 확률분포를 표현하는데 있어서 X가 가질 수 있는 모든 값(x)을 나열하며 확률을 대응시키기는 어렵다. 이 경우에는 확률변수 X가 가질 수 있는 특정 구간에서 확률이 어떻게 분포하는가를 나타낼 수 있는 함수를 이용하게 된다. 즉, 연속확률변수 X의 확률분포는 확률의 밀도를 나타내는 확률밀도함수에 의해 결정된다.

■ 확률밀도함수(probability density function)

다음을 만족하는 함수 $f(x)$를 연속확률변수 X의 확률밀도함수라 한다.

(i) 모든 x에 대해 $f(x) \geq 0$

(ii) $P(a \leq X \leq b) = \int_a^b f(x)dx$

(iii) $P(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} f(x)dx = 1$

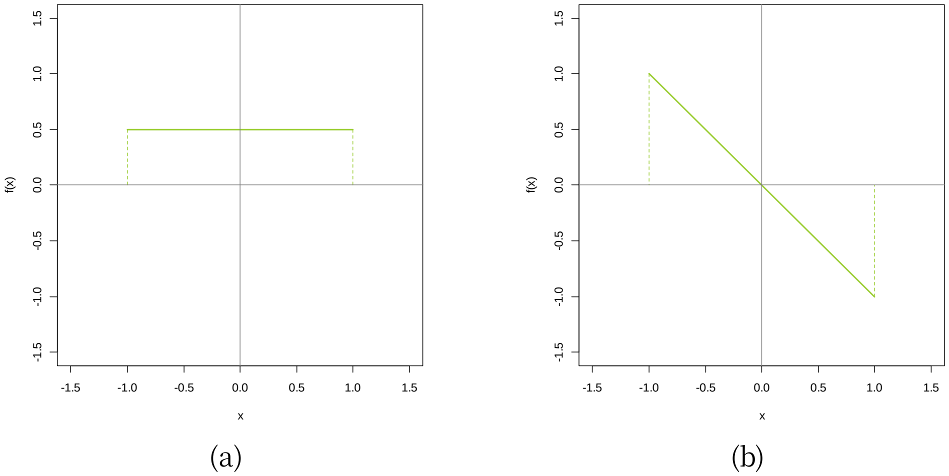
다음 [그림 4.1]을 통해 확률밀도함수를 구분해 보자. [그림 4.1] (a)의 경우에는 [-1, 1]의 구간에 대해 $f(x) = 0.5 > 0$인 직선이 표현되어 있다. 따라서

$$P(-1 \leq X \leq 1) = \int_{-1}^{1} f(x)dx = 1$$

이 된다. 또한 나머지 구간에 대해서는 $f(x) = 0$이므로

$$P(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} f(x)dx = 1$$

의 조건도 만족하게 된다. 즉, [그림 4.1]의 (a)에 표현된 $-1 \le x \le 1$에 대한 $f(x)=0.5$의 직선은 확률밀도함수이다.



[그림 4.1] 확률밀도함수의 구분

[그림 4.1] (b)의 경우에는 [-1, 1]의 구간에 대해 $f(x)=-x$의 직선이 표현되어 있다. 이 경우 해당 구간에 대해서는

$$P(-1 \le X \le 1) = \int_{-1}^{1} f(x)\,dx = 1$$

이 되며, 나머지 구간에 대해서는 $f(x)=0$이므로

$$P(-\infty \le X \le \infty) = \int_{-\infty}^{\infty} f(x)\,dx = 1$$

의 조건도 만족하게 된다. 그러나 $0 \le x \le 1$의 구간에 대해서 $f(x) \le 0$이므로 [그림 4.1]의 (b)에 표현된 $-1 \le x \le 1$에 대한 $f(x)=-x$의 직선은 확률밀도함수가 아니다.

[그림 4.1]을 토대로 정리를 해보면, 연속확률변수 X가 특정한 값 x를 갖게 되는 확률은 0임을 알 수 있다. 그러므로 어떤 구간의 확률을 구할 때는 그 구간의 경계점이 포함되는가 포함되지 않는가에 영향을 받지 않는다. 즉, 임의의 구간 $[a, b]$에 대한 확률은

$$P(a \le X \le b) = P(a < X \le b) = P(a \le X < b) = P(a < X < b)$$

가 된다.

연속확률변수의 기대값과 분산은

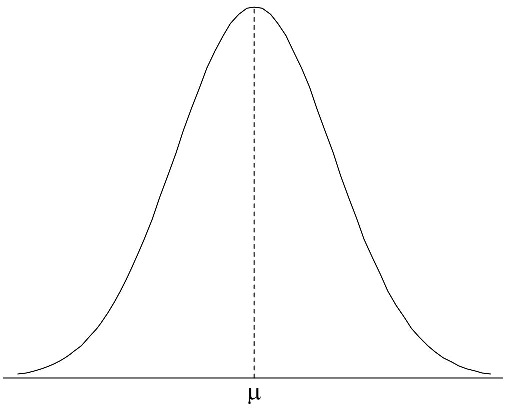
$$E(X) = \int x f(x) dx = \mu$$

$$Var(X) = \int (x - \mu)^2 f(x) dx = \sigma^2$$

와 같이 계산할 수 있다. 이를 이산확률변수의 경우와 비교해 보면, 이산확률변수의 경우에는 특정한 값 x를 갖게 되는 확률을 각각 구할 수 있기 때문에 각각의 값들을 다 더한 것이고, 연속확률변수의 경우에는 구간에 대한 확률로만 표현할 수 있기 때문에 적분의 계산방법을 응용하는 것일 뿐 기대값과 분산에 대한 의미상의 차이는 없음을 알 수 있다. 따라서 3.4절의 기대값과 분산의 성질은 연속확률변수에 대해서도 여전히 성립된다.

4.2 정규분포

통계학의 모든 분야에서 가장 중요하게 생각되는 대표적인 연속확률분포는 정규분포(normal distribution)이다. 정규곡선이라 불리는 정규분포의 확률밀도함수는 [그림 4.2]의 그래프와 같이 좌우대칭의 종모양(bell-shape) 곡선으로, 이는 자연현상과 산업현장 등의 여러 연구 분야에서 얻어지는 자료의 분포 형태를 묘사한다. 정규분포는 가우스(Gauss, K.F.)가 같은 양을 반복 측정하는 과정에서 생긴 오차들을 연구하다 발견하였기에, 그의 이름을 따서 종종 가우스 분포(Gaussian distribution)라고도 하며, 1733년에 드 므와브로(De Moivre)는 정규곡선의 수학적 방정식을 유도한 바 있다.



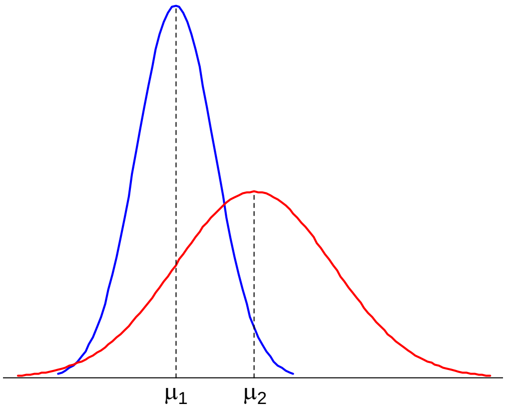
[그림 4.2] 정규곡선

[그림 4.2]의 종 모양의 확률밀도함수를 가지는 연속형확률변수 X를 정규확률변수(normal random variable)라 한다. 정규확률변수의 확률분포에 대한 수학적 방정식은 모집단에 대한 평균 μ와 분산 σ^2에 의존한다. 그래서 기호로

$$X \sim N(\mu, \sigma^2)$$

와 같이 나타낸다.

정규분포에서 μ와 σ^2이 가지는 의미를 살펴보면, μ는 [그림 4.2]에서와 같이 확률분포의 중심위치를 나타내고 σ^2은 정규곡선이 퍼져 있는 정도를 나타낸다.



[그림 4.3] μ와 σ^2에 따른 정규곡선 비교

[그림 4.3]에는 두 개의 정규곡선이 그려져 있다. 좌측 푸른색의 정규곡선은 $N(\mu_{1,}\sigma_1^2)$을 나타내고, 우측 붉은색의 정규곡선은 $N(\mu_{2,}\sigma_2^2)$을 나타낸다. 이때, $\mu_1 < \mu_2$이고 $\sigma_1^2 < \sigma_2^2$이다. 즉, μ_1보다 μ_2가 크기 때문에 붉은색의 정규곡선의 중심위치는 푸른색의 정규곡선보다 우측에 위치하고 있으며, σ_1^2이 σ_2^2보다 작기 때문에 붉은색의 정규곡선보다 푸른색의 정규곡선이 좁고 뾰족하게 나타난다.

4.3 정규분포의 확률계산

μ와 σ^2에 따라 다양한 모양으로 나타나는 정규분포 중에서 평균이 0이고 분산이 1인 정규분포를 표준정규분포(standard normal distribution)라 한다. 일반적으로 표준정규분포를 따르는 확률변수는 Z로 표현하는데, $Z \sim N(0,1)$이므로 확률변수는 Z는 $E(Z) = \mu_z = 0$을 중심으로 좌우대칭인 정규분포를 따르며, 분산은 $Var(Z) = \sigma_z^2 = 1$와 같다.

정규분포의 유용한 특징 중 하나는 정규분포를 따르는 확률변수에 상수를 이용한 사칙연산을 하더라도 여전히 정규분포를 따르게 된다는 것이다. 다시 말해, 만약 $X \sim N(\mu, \sigma^2)$이라고 하면 확률변수 X에 상수를 이용한 사칙연산을 적용한 새로운 확률변수는 평균과 분산만 바뀔 뿐 여전히 정규분포를 따르게 된다는 것이다. 이와 같은 성질을 이용하여 일반적인 정규분포를 따르는 확률변수 X에 대한 확률을 표준정규분포를 따르는 확률변수 Z에 대한 확률로 표현 가능하게 된다. 이처럼 확률변수 X를 평균이 0이고 분산이 1인 확률변수 Z로 변환하는 과정을 표준화변환이라고 한다. 확률변수 X의 기대값을 μ_x, 분산을 σ_x^2이라 하면, 확률변수 X에 대한 표준화변환의 방법은

$$\frac{X - \mu_x}{\sigma_x}$$

와 같으며, 표준화변수를 Z라 하면 Z의 기대값과 분산은 다음과 같다.

$$E(Z) = E\left[\frac{X - \mu_x}{\sigma_x}\right] = \frac{1}{\sigma_x} E[X - \mu_x] = \frac{1}{\sigma_x}(\mu_x - \mu_x) = 0$$

$$Var(Z) = Var\left[\frac{X - \mu_x}{\sigma_x}\right] = \frac{1}{\sigma_x^2} Var(X) = \frac{\sigma_x}{\sigma_x} = 1 .$$

따라서 $X \sim N(\mu, \sigma^2)$이라고 하면 표준화변수 Z는 평균이 0이고 분산이 1인 표준정규분포를 따르게 된다.

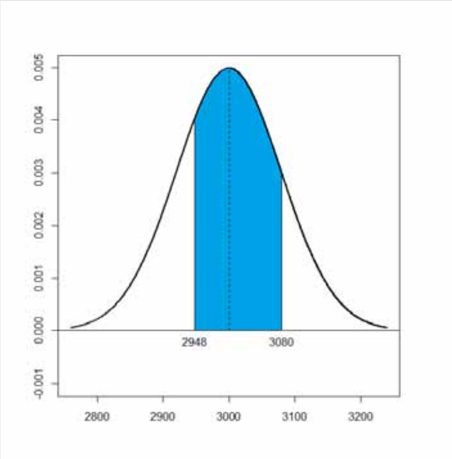
보기 4.1

어떤 회사에서 제조되는 전구의 수명시간은 정규분포를 따른다고 한다. 제조되는 전구들에 대한 수명시간에 대한 평균은 3,000시간이고 표준편차는 80시간이라 할 때, 임의로 선택한 전구 1개의 수명시간이 2,948시간에서 3,080시간 사이일 확률을 구하여라.

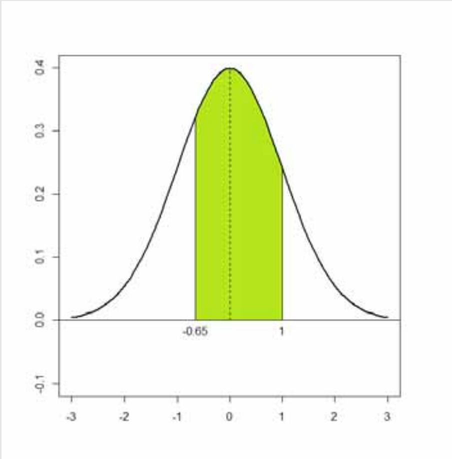
확률변수 X는 이 회사에서 제조되는 전구의 수명시간이다. 이러한 확률변수 X는 평균이 3,000시간, 표준편차가 80시간인 정규분포를 따른다고 하였으므로, $X \sim N(3000, 80^2)$이다. 이에 임의로 선택한 전구 1개의 수명시간이 2,948시간에서 3,080시간 사이일 확률은

$$\begin{aligned} P(2948 < X < 3080) &= P\left(\frac{2948-3000}{80} < \frac{X-3000}{80} < \frac{3080-3000}{80}\right) \\ &= P(-0.65 < Z < 1.00) \end{aligned}$$

와 같이 표현할 수 있다. 이들 구간을 그림으로 표현하면 아래 [그림 4.4]와 같이 동일한 형태임을 알 수 있다.



(a) $P(2948 < X < 3080)$



(b) $P(-0.65 < Z < 1.00)$

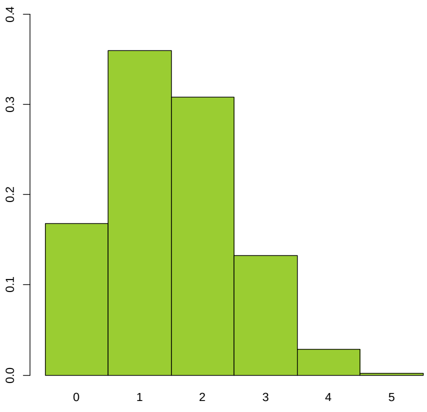
[그림 4.4] 정규분포와 표준정규분포의 확률

따라서 구하고자 하는 확률의 값은 부록의 표를 이용하여 다음과 같이 계산되어진다.

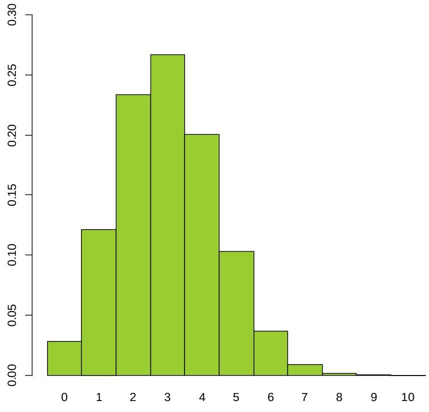
$$
\begin{aligned}
P(-0.65 < Z < 1.00) &= P(Z < 1.00) - P(Z < -0.65) \\
&= 0.8413 - 0.2578 \\
&= 0.5835
\end{aligned}
$$

4.4 이항분포의 정규근사

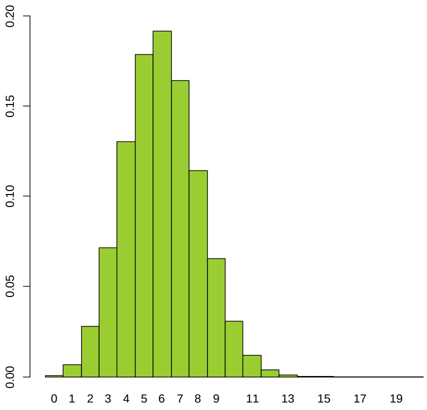
3장에서 우리는 이항분포는 성공확률이 p인 베르누이 시행을 n번 반복한 경우 성공 횟수가 가지는 확률분포임을 익힌 바 있다. 이항분포의 특징은 n이 증가함에 따라 분포의 형태가 점차 대칭에 가까워지고 종 모양을 이루게 된다.



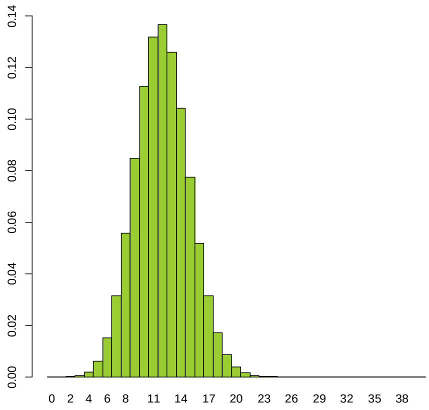
(a) $p = 0.4,\ n = 5$



(b) $p = 0.4,\ n = 10$



(c) $p = 0.4,\ n = 20$



(d) $p = 0.4,\ n = 40$

[그림 4.5] 이항분포의 확률 히스토그램

[그림 4.5]는 $p=0.3$ 이고 $n=5, 10, 20, 40$ 인 경우에 대한 이항분포의 확률 히스토그램이다. 이 그림을 통해 n 이 증가함에 따라 분포의 형태가 점차 좌우대칭의 종 모양에 가까워짐을 확인할 수 있다.

이와 같은 경향은 n 이 커짐에 따라 더욱 확실하게 드러나게 되는데, 실제로 이항분포는 n 이 증가함에 따라 근사적으로 정규분포를 따르게 된다. 이때, 정규분포의 평균과 분산은 이항분포에서와 일치하여야 하므로 다음과 같다.

$$\mu=np\,,\quad \sigma^2=np(1-p)$$

그런데 여기서 n 이 어느 정도 커져야 이항분포를 따르는 확률변수가 근사적으로 정규분포를 따르게 될까? 이에 대한 명확한 답은 존재하지 않지만 일반적으로 np 나 $n(1-p)$ 가 모두 10 이상이면 이항분포 상에서의 확률과 근사정규분포 상에서의 확률적 차이가 아주 적다고 알려져 있다.

■ 이항분포의 정규근사

확률변수 X가 성공 횟수를 나타내는 이산형 확률변수이고 $X \sim \mathrm{Bin}\,(n, p\,)$ 일 때, np 나 $n(1-p)$ 가 모두 충분히 클 경우(보통 10 이상)에 확률변수 X는 근사적으로 다음의 정규분포를 따르게 된다.

$$X \sim N(np, np(1-p)\,)$$

보기 4.2

확률변수 X가 $n=150$, $p=0.6$ 의 이항분포를 따른다고 하자. 이러한 확률변수 X가 82 이상 102 미만일 확률은 얼마인가?.

우선 $X \sim \mathrm{Bin}\,(150, 0.6\,)$ 이라고 했으므로, 확률변수 X는 이산형 확률변수이다. 따라서 X가 82 이상 102 미만일 확률은

$$P(82 \le X < 102) = P(X=82) + \cdots + P(X=101)$$

$$= {}_{150}\mathrm{C}_{82} \times (0.6)^{82} \times (0.4)^{68} + \cdots + {}_{150}\mathrm{C}_{101} \times (0.6)^{101} \times (0.4)^{49}$$

와 같이 계산하여야만 정확한 확률을 계산할 수 있다. 그러나 이와 같은 계산을 수작업으로 진행하기에는 많은 계산량과 시간을 요구하게 된다.

이러한 경우 조건만 충족된다면 이항분포의 정규근사를 통해 근사확률을 계산할 수 있다. 이 경우 $np=90$ 이고 $n(1-p)=60$ 이므로 둘 다 10 이상이다. 따라서 확률변수 X는 평균 $\mu=np=90$ 이고 분산 $\sigma^2=np(1-p)=36$ 인 근사정규분포를 따른다고 할 수 있다.

이제 X가 82 이상 102 미만일 확률을 계산하여야 하는데, 이때 주의하여야 할 점이 있다. 이항분포는 이산형 확률분포로서 [그림 4.5]에서 확인할 수 있듯이 확률분포가 계단형으로 나타난다. 그러나 정규분포는 연속형 확률분포로 확률밀도함수의 형태가 좌우대칭의 종 모양 곡선으로 나타난다. 따라서 이산형 확률분포를 따르는 확률변수 X에 대해 연속형 확률분포로의 근사를 위한 연속성 수정(continuity correction)을 해야만 한다. 연속성 수정 방법은 이항분포에서 $[a \le X \le b]$ 일 확률을 정규분포에서 $(a-0.5 < X < b+0.5)$ 일 확률로 근사시키는 것이다. 즉, 이항분포 상에서

$$P(82 \le X < 102) = P(82 \le X \le 101)$$

이므로, 이에 대해 연속성 수정을 하게 되면 정규분포 상에서는

$$P(81.5 < X < 101.5)$$

의 확률을 계산하여야 한다. 그러므로 다음과 같이 계산할 수 있다.

$$P(81.5 < X < 101.5) = P\left(\frac{81.5-90}{\sqrt{36}} < \frac{X-90}{\sqrt{36}} < \frac{101.5-90}{\sqrt{36}}\right)$$
$$= P(-1.42 < Z < 1.92) = 0.8948$$

4.5 R-프로그램 실습

[프로그램 4.1]은 [보기 4.1]의 전구의 수명시간에 대해 정규분포를 활용한 R-프로그램을 보여주고 있다. R-프로그램의 내장함수 중 pnorm을 이용하면 정규분포에서의 확률을 쉽게 계산할 수 있다. 구문형식은

```
pnorom(x, mean = 0, sd = 1, lower.tail = TRUE)
```

이며, 여기서 x는 확률변수 X의 특정값을 의미한다. 그리고 mean은 평균(μ)을, sd는 표준편차(σ)를 의미하며, 이들 값을 입력하지 않으면 표준정규분포에서의 확률을 계산해준다. 그리고 lower.tail은 확률계산의 방향을 선택하는 옵션으로 그 값을 TRUE(또는 T)로 설정하면 $P(X \leq x)$가 계산되고, FALSE(또는 F)로 하면 $P(X \geq x)$가 계산되며, 기본값은 TRUE이므로 TRUE(또는 T)로 설정 시에는 생략 가능하다.

[프로그램 4.1] [보기 4.1]의 전구의 수명시간에 대한 확률

```
mu = 3000
sigma = 80
x1 = 2948
x2 = 3080
pnorm(x2,mean=mu,sd=sigma) - pnorm(x1,mean=mu,sd=sigma)

z1 = (x1-mu)/sigma
z2 = (x2-mu)/sigma
pnorm(z2) - pnorm(z1)
```

[결과 4.1] [프로그램 4.1]의 수행 결과

```
0.5834986
0.5834986
```

[프로그램 4.2]는 [보기 4.2]의 이항분포의 정규근사 확률을 활용한 R-프로그램을 보여주고 있다. pbinom 함수를 이용하여 이항분포에서 직접적으로 계산한 확률(0.8945357)과 pnorm 함수를 이용한 정규근사 확률(0.8940697)이 거의 비슷한 값을 가짐을 확인할 수 있다. pbinom 함수의 이용방법은 3.6절을 참고하기 바란다.

[프로그램 4.2] [보기 4.2]의 이항분포에서의 확률과 정규근사 확률

```
n = 150
p = 0.6
pbinom(101,n,p) - pbinom(81,n,p)

mu = n*p
sigma = sqrt(n*p*(1-p))
pnorm(101.5,mu,sigma) - pnorm(81.5,mu,sigma)
```

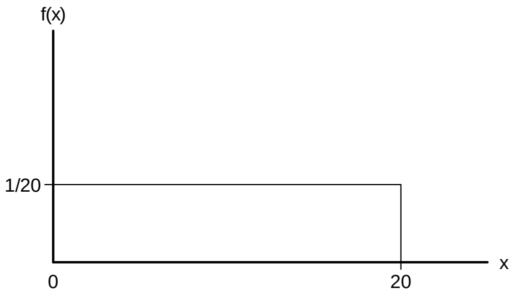
[결과 4.2] [프로그램 4.2]의 수행 결과

```
0.8945357
0.8940697
```

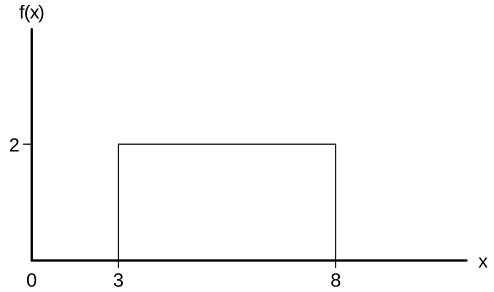
4.6 연습문제

4.1 다음의 그림이 연속확률변수의 확률밀도함수라 할 수 있는가?

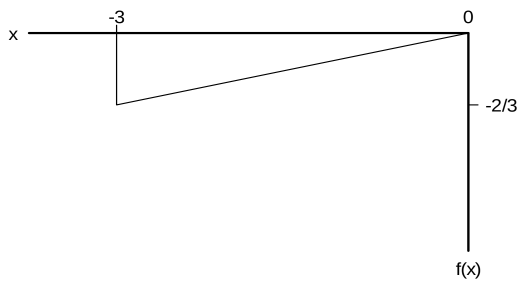
(1)



(2)



(3)



4.2 확률변수 X가 다음의 정규분포를 따른다고 가정하자. 표준화변수 Z가 주어질 경우, 확률변수 X의 값은 얼마인가?

(1) $X \sim N(2, 6^2),\ Z = 3$

(2) $X \sim N(8, 1^2),\ Z = -2.25$

(3) $X \sim N(9, 5^2),\ Z = -0.5$

4.3 외과 수술을 받은 환자의 회복 기간은 평균이 5.3일, 표준편차가 2.1일인 정규분포를 따른다고 한다.

(1) 환자의 회복 기간이 2일 이상 걸릴 확률은 얼마인가?

(2) 환자의 회복 기간이 5일에서 10일 사이일 확률은 얼마인가?

4.4 IQ는 평균이 100이고 표준편차가 15인 정규분포를 따른다고 할 때, 다음의 물음에 답하라.

(1) IQ가 120보다 클 확률은 얼마인가?

(2) 멘사는 IQ가 상위 2%인 사람들의 모임이라고 하자. IQ가 최소 몇 이상이어야 멘사 회원이 될 수 있는가?

4.5 초등학생의 70%가 스마트폰을 소지하고 있다고 한다. 임의로 선택된 초등학생 200명 중 스마트폰을 소지하고 있는 학생의 수가 125명 이상 155명 이하일 확률은 얼마인가?

4.6 미국 사람들이 하루에 소모하는 지방 칼로리의 비율(단위:%)은 평균 36, 표준편차 10인 정규분포를 따른다고 한다.

(1) 임의로 선택된 사람의 지방 칼로리의 소모 비율이 25일 경우, 표준화변수 Z는 얼마인가?

(2) 임의로 선택된 사람의 지방 칼로리의 소모 비율이 40 이상일 확률은 얼마인가?

(3) 지방 칼로리의 소모 비율이 하위 25%에 포함되려면 최대 몇 이하이어야 하는가?

4.7 한 야구선수가 경기에서 안타를 칠 확률이 0.25라고 알려져 있다. 이 선수는 올해 총 300타석을 섰다고 한다.

(1) 안타를 친 횟수가 50회 미만일 확률은 얼마인가?

(2) 안타를 친 횟수가 100회 이상일 확률은 얼마인가?

4.8 앞면이 나올 확률이 20%인 구부러진 동전을 총 100번 던졌다고 한다.

(1) 앞면이 10번 이하로 나올 확률은 얼마인가?

(2) 앞면이 13번 초과 24번 미만으로 나올 확률은 얼마인가?

5장

표집분포와 중심극한정리

R과 함께하는

통계학의 이해

5.1 표집분포

우리는 모집단의 특성 중 어떠한 수치, 예를 들면, 모평균이나 모비율 또는 모분산 등에 관심이 있는 경우가 많이 있다. 이러한 모집단에 대한 수치적 특성값을 모수(parameter)라고 하며, 통계학에서 주요 관심대상이 된다. 모수의 참값을 알려면 모집단 전체를 조사해야만 한다. 그러나 모집단 전체를 조사한다는 것은 불가능하거나 가능하더라도 많은 시간과 비용을 요하는 경우가 허다하다. 이러한 이유로 우리는 일반적으로 모수의 참값을 정확하게 알아낼 수가 없다. 그렇기 때문에 우리는 모집단으로부터 적절한 크기의 표본을 추출하게 되고, 해당 표본으로부터 획득한 정보를 이용하여 모수에 대한 추론을 하게 된다. 이러한 표본으로부터 획득한 수치적 정보를 통계량(statistic)이라고 부른다. 예를 들어 표본평균은 표본의 관측값에 의해 계산되어진 수치적 정보이므로 통계량이다. 이러한 통계량들을 활용할 때 유의하여야 할 점은 다음과 같다.

(i) 표본은 모집단의 일부이므로, 표본으로부터 획득한 통계량의 값은 모수의 참값과 일반적으로 같지 않다.
(ii) 통계량의 값은 추출된 표본에 영향을 받으므로, 다른 표본을 추출하게 되면 그 값이 변한다.

예를 들어, 어느 제조업체에서 자사제품에 대한 고객들의 만족도를 알아보고자 한다고 가정하자. 이 경우 실제 알고자 하는 주요 관심대상은 해당 제품을 구매한 전체 고객들의 평균 만족도이다. 즉, 모집단은 해당 제품을 구매한 고객 전체이며, 모수는 자사 제품에 대한 고객 전체의 평균만족도(μ)이다. 고객 전체의 평균만족도(μ)의 참값은 미지의 상수이다. 이 값을 알기 위해 고객 전체를 조사하는 것은 불가능하므로, 임의로 선정한 100명의 고객들을 표본으로 선정하여 5점 만점의 설문조사를 실시하였고, 조사 결과 평균이 $\overline{x} = 3.75$점이었다고 하자. 이렇게 구한 평균은 표본평균이다. 이 값은 궁극적으로 알고자 하는 고객 전체의 평균만족도(μ)의 참값과 일치하지는 않을 것이다. 또한, 또 다른 100명의 고객을 표본으로 선정하여 재조사를 실시한다면, 이들 100명의 평균만족도 값이 이전 조사 결과의 3.75점으로 나올 가능성은 매우 희박하다. 그러나 우리는 재조사 결과가 3.75점과 비슷한 값을 가질 것이라고 생각할 수 있다. 즉, 통계량의 값은 표본을 추출할 때마다 어떠한 값을 중심으로 약간의 변동을 가지게 된다는 것이다. 이러한 변동을 파악할 수만 있다면 현재의 표본평균이 실제 모평균

의 참값과 얼마나 가까운지를 알 수 있을 것이다.

통계량은 그 자체가 하나의 확률변수로서 확률분포를 가지게 된다. 이처럼 통계량이 가지는 확률분포를 표집분포(sampling distribution)라고 한다. 반복되는 표본 추출과정에서 발생하는 통계량의 값이 가지는 변동은 이 표집분포에 의해 설명될 수 있다. 표집분포는 모집단의 분포에 영향을 받기도 하고, 표본의 크기 n에도 영향을 받는다. 다음 보기를 통해 표본평균의 확률분포를 살펴보기로 하자.

보기 5.1

하나의 공정한 주사위를 던져 윗면에 나타난 수가 1이나 2이면 1점을, 3이나 4이면 3점을, 5나 6이면 5점을 획득할 수 있는 게임을 생각해보자. 주사위를 두 번 던졌을 때 획득할 수 있는 평균점수의 확률분포를 구하라.

우선 첫 번째 주사위를 던졌을 때 획득할 수 있는 점수를 확률변수 X_1이라고 하면, 확률변수 X_1가 가질 수 있는 값과 확률분포는 다음과 같다.

x_1	1	3	5
$f(x_1)$	1/3	1/3	1/3

다음으로 두 번째 주사위를 던졌을 때 획득할 수 있는 점수를 확률변수 X_2라고 하면, 확률변수 X_2가 가질 수 있는 값과 그때의 확률분포는 확률변수 X_1의 경우와 같다.

이제, 주사위를 두 번 던졌을 때 획득할 수 있는 평균점수를 생각해 보자. 이 경우의 평균점수는 확률변수 X_1과 X_2가 가지는 값의 평균이다. 즉,

$$\frac{1}{2}(x_1 + x_2)$$

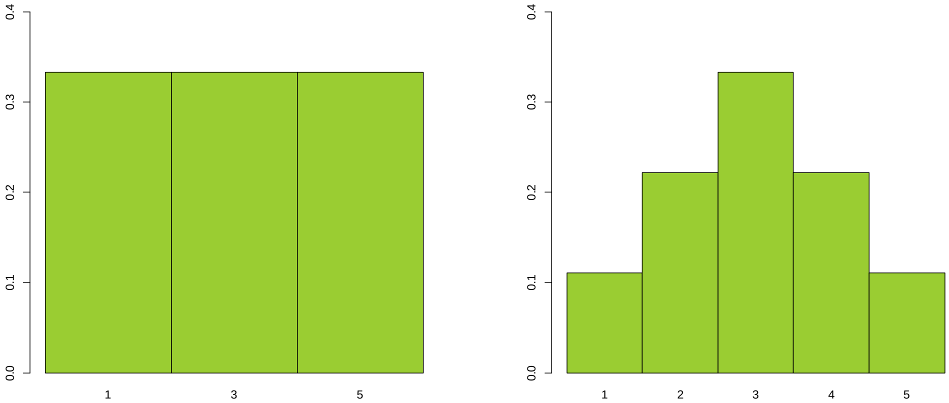
이므로, 표본평균의 값을 의미한다. 따라서 주사위를 두 번 던졌을 때 획득할 수 있는 평균점수를 확률변수 $\overline{X} = (X_1 + X_2)/2$라고 하면, 확률변수 X_1와 X_2가 가지는 값과 그에 대응되는 확률변수 $\overline{X}$가 가질 수 있는 값은 다음과 같이 정리할 수 있다.

(x_1, x_2)	(1,1)	(1,3)	(1,5)	(3,1)	(3,3)	(3,5)	(5,1)	(5,3)	(5,5)
$\overline{x}$	1	2	3	2	3	4	3	4	5

따라서 확률변수 $\overline{X}$가 가질 수 있는 값과 그때의 확률분포는 다음과 같다.

$\overline{x}$	1	2	3	4	5
$f(\overline{x})$	1/9	2/9	3/9	2/9	1/9

일반적으로 크기가 큰 모집단으로부터 임의 추출된 크기 n의 표본 $X_1, X_2, \cdots, X_n$은 서로 독립이고 모두 모집단의 분포와 같은 분포를 갖는 것으로 간주한다. 이러한 $X_1, X_2, \cdots, X_n$을 확률표본(random sample)이라고 부른다. 위 보기에서 X_1과 X_2의 분포는 모집단의 분포와 동일하다. 즉, X_1과 X_2는 크기 2의 확률표본이다. 그러나 이러한 확률표본으로부터 구한 표본평균 $\overline{X}$의 확률분포는 위 보기와 [그림 5.1]에서 확인할 수 있듯이 모집단의 확률분포보다 중심으로 더 몰리게 된다.



(a) X의 확률분포 (b) $\overline{X}$의 확률분포

[그림 5.1] 확률밀도함수의 구분

5.2 표본평균의 분포와 중심극한정리

평균이 μ 이고 분산이 σ^2 인 모집단으로부터 크기가 n 인 표본 $X_1, X_2, \cdots, X_n$ 을 추출한 경우, 표본평균 $\overline{X}$ 의 기대값, 분산과 표준편차는 다음과 같다.

■ 표본평균 $\overline{X}$ 의 기대값

크기가 n 인 확률표본 $X_1, X_2, \cdots, X_n$ 에 대해

$$E(X_i) = \mu, \ Var(X_i) = \sigma^2, \ i = 1, \cdots, n$$

이고,

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

일 때, 표본평균 $\overline{X}$ 의 기대값과 분산, 표준편차는 다음과 같다.

$$E(\overline{X}) = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\left[\sum_{i=1}^{n} E(X_i)\right] = \frac{1}{n}(\mu + \cdots + \mu) = \mu$$

$$Var(\overline{X}) = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\left[\sum_{i=1}^{n} Var(X_i)\right] = \frac{1}{n^2}(\sigma^2 + \cdots + \sigma^2) = \frac{\sigma^2}{n}$$

$$\sqrt{Var(\overline{X})} = \frac{\sigma}{\sqrt{n}}$$

여기서 표본평균이 가지는 분포의 중심은 언제나 모집단의 중심인 모평균 μ와 일치한다. 그리고 표본평균의 값에 대한 변동을 측정하는 분산과 표준편차는 각각 모집단의 분산과 표준편차에 영향을 받을 뿐만 아니라 표본의 크기에도 영향을 받는다. 즉, 표본의 크기 n 이 증가함에 따라 표본평균의 분산과 표준편차는 감소하게 되어 표본평균의 분포가 모평균 μ 를 중심으로 더욱 집중됨을 의미한다.

이제 표본평균의 기대값과 분산에 대한 특성은 알게 되었는데, 표본평균이 가지는 확률분포는 어떠할까? 우선, 모집단의 확률분포가 정규분포라면 표본평균의 확률분포는 다음과 같은 특성을 지니게 된다.

■ 정규모집단으로부터의 표본평균에 대한 확률분포

크기가 n인 확률표본 $X_1, X_2, \cdots, X_n$이

$$X_i \sim N(\mu, \sigma^2), \quad i = 1, \cdots, n$$

이면, 표본평균 $\overline{X}$의 확률분포는 다음과 같은 정규분포를 따르게 된다.

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

만약 모집단의 확률분포가 정규분포가 아닌 경우에는, 표본평균 $\overline{X}$의 정확한 확률분포는 모집단의 확률분포에 따라 다르게 나타난다. 그러나 표본의 크기 n이 큰 경우에는 표본평균 $\overline{X}$의 확률분포가 모집단의 분포와 무관하게 중심극한정리에 의해서 근사적으로 정규분포를 따르게 된다.

■ 중심극한정리

평균이 μ이고 분산이 σ^2인 모집단으로부터 추출한 크기 n의 확률표본의 표본평균 $\overline{X}$는 표본의 크기가 큰 경우 (보통 30 이상), 근사적으로 평균이 μ이고 분산이 σ^2/n인 정규분포를 따르게 된다.

중심극한정리는 모집단의 확률분포가 연속형이거나 이산형, 혹은 대칭이거나 비대칭에 상관없이 표본의 크기가 충분히 크다면 표본평균의 확률분포가 근사적으로 정규분포를 따르게 된다는 것이다.

보기 5.2

만 2세 유아들의 신장은 평균이 87.6㎝이고 표준편차가 3.3㎝인 정규분포를 따른다고 한다. 만약 6명의 2세 유아들에 대해 신장을 조사하였을 경우, 이들의 평균 신장이 86.6㎝에서 89.4㎝ 사이일 확률은 얼마인가?

우선 모집단은 만 2세 유아들이고, 이들의 신장을 확률변수 X라 하면, 확률변수 X는 다음과 같은 정규분포를 따른다고 하였다.

$$X \sim N(87.6, 3.3^2)$$

따라서 이들 모집단으로부터 추출한 6명의 유아들에 대한 평균 신장을 확률변수 $\overline{X}$라고 하면, 확률변수 $\overline{X}$의 확률분포는 모집단이 정규분포를 따르므로 다음과 같이 정의된다.

$$\overline{X} \sim N(87.6, 3.3^2/6)$$

그러므로 평균 신장이 86.6㎝에서 89.4㎝ 사이일 확률은

$$\begin{aligned} P(86.6 < \overline{X} < 89.4) &= P\left(\frac{86.6-87.6}{3.3/\sqrt{6}} < \frac{\overline{X}-87.6}{3.3/\sqrt{6}} < \frac{89.4-87.6}{3.3/\sqrt{6}}\right) \\ &= P(-0.74 < Z < 1.34) = 0.6802 \end{aligned}$$

보기 5.3

어떤 열매의 무게는 평균이 0.5g이고 표준편차가 0.15g이라고 한다. 100개의 열매를 임의로 선택한 경우, 이들의 평균 무게가 0.48g에서 0.53g 사이일 확률은 얼마인가?

우선 모집단은 어떤 특정 열매들이고, 이들의 무게를 확률변수 X라 하면, 확률변수 X는 다음과 같은 특징을 가지고 있다.

$$E(X) = 0.5,\ Var(X) = 0.15^2$$

이들 모집단으로부터 추출한 100개의 열매들에 대한 평균 무게를 확률변수 $\overline{X}$ 라고 하면, 모집단의 확률분포는 모른다고 하더라도 표본의 크기가 충분히 크므로 중심극한정리에 의해 확률변수 $\overline{X}$ 의 확률분포는 다음과 같은 정규분포로 정의된다.

$$\overline{X} \sim N(0.5, 0.15^2/100)$$

그러므로 평균 무게가 0.48g에서 0.53g 사이일 확률은

$$\begin{aligned} P(0.48 < \overline{X} < 0.53) &= P\left(\frac{0.48-0.5}{0.15/\sqrt{100}} < \frac{\overline{X}-0.5}{0.15/\sqrt{100}} < \frac{0.53-0.5}{0.15/\sqrt{100}}\right) \\ &= P(-1.33 < Z < 2.00) = 0.8854 \end{aligned}$$

5.3 R - 프로그램 실습

[프로그램 5.1]은 [보기 5.2]의 유아들의 신장에 대해 정규분포를 활용한 R - 프로그램을 보여주고 있다. R - 프로그램의 내장함수 중 pnorm을 이용하면 정규분포에서의 확률을 쉽게 계산할 수 있으며, 구문형식은 4.5절을 참고하기 바란다.

[프로그램 5.1] [보기 5.2]의 유아들의 신장에 대한 확률

```
mu = 87.6
sigma = 3.3
n = 6
x1 = 86.6
x2 = 89.4
pnorm(x2,mu,sigma/sqrt(n)) - pnorm(x1,mu,sigma/sqrt(n))

z1 = (x1-mu)/(sigma/sqrt(n))
z2 = (x2-mu)/(sigma/sqrt(n))
pnorm(z2) - pnorm(z1)
```

[결과 5.1] [프로그램 5.1]의 수행 결과

```
0.6802773
0.6802773
```

[프로그램 5.2]는 [보기 5.3]의 열매의 무게에 대해 정규분포를 활용한 R - 프로그램을 보여주고 있다.

[프로그램 5.2] [보기 5.3]의 열매의 무게에 대한 확률

```
mu = 0.5
sigma = 0.15
n = 100
x1 = 0.48
x2 = 0.53
pnorm(x2,mu,sigma/sqrt(n)) - pnorm(x1,mu,sigma/sqrt(n))

z1 = (x1-mu)/(sigma/sqrt(n))
z2 = (x2-mu)/(sigma/sqrt(n))
pnorm(z2) - pnorm(z1)
```

[결과 5.2] [프로그램 5.2]의 수행 결과

```
0.8860386
0.8860386
```

5.4 연습문제

5.1 통계학과 학생들의 형제 수를 조사하였더니, 1, 2, 3으로 이루어져 있었다. 두 명의 학생을 선택하였을 때 X_1, X_2를 각각 첫 번째와 두 번째 학생의 형제 수라고 하면, 표본평균 $\overline{X}$의 확률분포를 구하라.

5.2 연습문제 5.1에서 다음의 물음에 답하라.

(1) 확률변수 X의 평균 μ와 분산 σ^2을 구하라.

(2) 연습문제 5.1에서 구한 표본평균 $\overline{X}$의 확률분포표를 이용하여 $\overline{X}$의 기대값과 분산을 구하라.

(3) (2)에서 구한 값이 $E(\overline{X})=\mu$, $Var(\overline{X})=\dfrac{\sigma^2}{n}$와 같은지 확인하라.

5.3 경상도 지역의 휘발유 가격은 평균이 1,910원이고 표준편차가 220원이며, 모집단의 확률분포는 알려져 있지 않다. 경상도 지역에 위치한 주유소 60곳을 임의로 선택하여 휘발유 가격을 조사하였다. 주유소 60곳의 휘발유 가격의 평균은 어떠한 분포를 따르는가?

5.4 연습문제 5.3에서 주유소 60곳의 휘발유 가격의 평균이 1,850원에서 1,950원 사이에 있을 확률은 얼마인가?

5.5 1부터 10까지 적힌 숫자 카드가 있다. 10장의 카드 중 1장을 뽑아 숫자를 확인하고 넣은 후, 다시 10장의 카드 중 1장을 뽑는 것을 반복하는 게임을 한다고 하자. 카드를 뽑아 나온 수가 홀수이면 3점, 짝수이면 5점을 획득한다. 카드를 3번 반복하여 뽑았을 때 획득할 수 있는 평균 점수의 확률분포를 구하라.

5.6 야구 경기에서 외야로 날아간 공의 거리는 정규분포를 따르며 평균이 250피트, 표준편차가 50피트이다. 외야로 날아간 공 중 19개를 임의로 선택하였다.

(1) 19개 공의 날아간 거리의 평균과 표준편차는 얼마인가?

(2) 19개 공의 날아간 거리의 평균은 어떠한 분포를 따르는가?

(3) 19개 공의 날아간 거리가 240피트보다 작을 확률은 얼마인가?

5.7 한 시골에서 마을버스를 기다리는 시간은 평균이 30분이고 표준편차가 4분이며, 모집단의 확률분포는 알려져 있지 않다. 100명의 승객을 임의로 선택하여 버스를 기다린 시간을 조사하였다.

(1) 100명의 승객이 버스를 기다린 시간의 평균은 어떠한 분포를 따르는가?

(2) 100명의 승객이 버스를 기다린 시간의 평균이 29분에서 31분 사이에 있을 확률은 얼마인가?

5.8 한 커피전문점의 하루 매출은 정규분포를 따르며, 평균이 110만 원이고 표준편차가 12만 원이다. 임의로 선택된 날부터 일주일간 매출을 조사한 결과, 평균 매출이 120만 원에서 125만 원 사이에 있을 확률을 구하라.

5.9 지난 수학 모의고사 성적은 평균이 80점, 표준편차가 8점이라고 한다. 지난 모의고사를 응시한 학생 중 40명을 임의로 선택하여 점수를 조사한 결과, 평균 점수가 83점 이상일 확률을 구하라.

6장

추정

R과 함께하는

통계학의 이해

6.1 통계적 추론

통계적 추론(statistical inference)은 모집단의 수치적 특성을 나타내는 모수(parameter)에 대한 정보를 얻어내기 위한 일련의 과정을 말한다. 모수의 값은 모르는 것이 보통이다. 따라서 이를 알아보기 위해선 모집단으로부터 추출된 표본으로부터 모수와 관련된 통계량(statistic)들의 값을 계산하고 이것을 이용하여 모집단의 특성 즉, 모수를 알아보는 것이다. 예를 들자면, 어떤 회사에서 생산되는 전구의 평균수명을 μ라 하자. 여기서 모집단은 해당 회사에서 생산되는 모든 전구를 말하며, 평균수명 μ는 이 모집단에 대한 모수 중 하나인 전체 평균을 나타낸다. 평균수명 μ의 값을 정확히 알고자 한다면 모집단 전체 즉, 생산되는 전구 전체에 대해 수명시간을 측정하여야 하지만 전구 전체를 측정하는 것은 실제로 불가능하다. 따라서 평균수명 μ를 알기 위해서는 생산된 전구 중 30개의 표본을 임의로 수집하여, 이들 전구에 대해서만 실제 수명 시간을 측정하는 것이다. 이들 30개 전구의 평균을 구해 보았더니 1,950시간이었고 하자. 계산된 평균수명 1,950시간은 통계량 중 하나인 표본평균($\bar{x}$)으로 그 값은 모수인 모평균 μ와 비슷한 값을 가지게 될 것이다. 이와 같은 방법으로 통계량을 이용하여 모수에 대한 추론을 진행하게 된다.

통계적 추론에는 조사 또는 연구의 목적에 따라 일반적으로 추정(estimation)과 가설 검정(hypotheses testing)으로 나눌 수 있다. 모수의 참값은 미지의 상수로서 모집단 전부를 조사하지 않고는 정확히 알아낼 수 없기에 해당 모수에 대한 추측값을 얻되, 그 값의 정밀도를 같이 구하고자 하는 것을 추정이라고 한다. 그리고 표본의 자료가 모수의 참값에 대한 조사자의 추측을 뒷받침하는지 혹은 반증하는지를 결정하는 것을 가설 검정이라고 부른다. 이 장에서는 추정에 관한 내용을 다루고 7장에서 가설 검정에 관한 내용을 다루도록 하겠다.

6.2 모평균에 대한 점추정

모수의 추정에는 점추정과 구간추정이 있다. 점추정(point estimation)이란 모수의 참값과 유사할 것이라고 예상되는 하나의 값을 제시하는 것이며, 구간추정(interval estimation)은 모수의 참값을 포함할 것으로 예상되는 적절한 구간을 제시하는 것이다. 이 절에서는 우선 모평균에 대한 점추정에 관하여 알아보고자 한다.

모수를 추정하기 위해 모집단에서 크기가 n인 표본을 임의로 추출할 때 이를 n개의 확률변수 $X_1, X_2, \cdots, X_n$으로 표현한다. 점추정이란 추정하고자 하는 하나의 모수에 대하여 이들 n개의 확률변수를 이용하여 하나의 통계량을 만들고, 나아가 주어진 표본으로부터 그 실제값을 계산하여 하나의 수치를 제시하려고 하는 것이다. 이와 같은 방법으로 모수를 추정하기 위해 만들어진 통계량을 *추정량*(estimator)이라 하고, 주어진 표본으로부터 계산된 추정량의 실제값을 추정치(estimate)라고 한다. 예를 들어 모평균 μ를 추정하고자 할 때, 직관적으로 가장 타당한 추정량은 표본평균

$$\hat{\mu} = \overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

이 될 것이다. 여기서 $\hat{\mu}$는 모평균 μ에 대한 추정량을 의미한다.

이러한 추정량은 확률변수들로부터 만들어진 하나의 확률변수이므로 추출된 표본의 값에 따라 그 값 즉, 추정치가 달라질 수 있다. 이런 수치들의 변동은 추정량의 정확도와 관계가 있는데, 이 정확도를 측정하기 위해 추정량의 표준편차를 계산할 필요가 있다. 이를 표준오차(standard error, $S.E.$)라고 부른다. 모평균 μ에 대한 가장 적절한 추정량은 표본평균 $\overline{X}$임을 앞서 언급한 바 있다. 표본평균의 경우 표준오차는 다음과 같이 구할 수 있다.

$$E(\overline{X}) = \mu, \quad S.E.(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

여기서 σ는 모집단의 표준편차를 의미한다. 따라서 표본평균 $\overline{X}$를 이용하여 모평균 μ를 추정하고자 할 경우에는 표본의 크기 n이 클수록 표준오차가 작아져서 보다 정확한 추정이 가능해짐을 알 수 있다. 그러나 실제로 $S.E.(\overline{X})$의 값은 σ를 모르는 경우 계산할 수 없다. σ는 모집단의 표준편차이므로 하나의 모수이기 때문이다. 이러한 경우에는 σ를 표본표준편차

$$\hat{\sigma} = s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

를 이용하여 추정할 수 있다.

■ 모평균에 대한 점추정

모　수 : 모평균(μ)

표　본 : 평균이 μ이고 표준편차가 σ인 모집단에서 임의로 추출한 $X_1, X_2, \cdots, X_n$

추 정 량 : 표본평균($\overline{X}$)

표준오차 : $S.E.(\overline{X}) = \sigma/\sqrt{n}$, (표준오차의 추정량 : $s/\sqrt{n}$)

보기 6.1

어린 소나무의 성장을 연구하기 위하여 1년생 붉은 소나무 묘목 40그루의 크기를 조사한 자료가 다음과 같다. 해당 자료를 이용하여 전체 1년생 붉은 소나무 묘목의 평균 크기에 대한 추정치와 표준오차는 얼마인가?

2.6	1.9	1.8	1.6	1.4	2.2	1.2	1.6	1.6	1.5
1.4	1.6	2.3	1.5	1.1	1.6	2.0	1.5	1.7	1.5
1.6	2.1	2.8	1.0	1.2	1.2	1.8	1.7	0.8	1.5
2.0	2.2	1.5	1.6	2.2	2.1	3.1	1.7	1.7	1.2

여기서 모집단은 표본의 40개 묘목을 포함한 모든 1년생 붉은 소나무 묘목들이다. 모집단의 확률분포는 알려져 있지 않고 묘목의 크기에 대한 모평균 μ나 모분산 σ^2과 같은 모수들 역시 알려져 있지 않다. 하지만 제시된 40개의 관측값은 모집단으로부터 추출한 확률표본이므로, 모

평균 μ에 대하여 다음과 같이 점추정을 할 수 있을 것이다.

먼저, 모평균 μ를 확률표본으로부터 추정할 때, 가장 직관적인 추정량은 표본평균 $\overline{X}$ 이므로, 모평균 μ에 대한 추정치는

$$\overline{x} = \frac{1}{40} \times (2.6 + 1.9 + \cdots + 1.7 + 1.2) = 1.715$$

이다. 다음으로 표준오차는 $S.E.(\overline{X}) = \sigma / \sqrt{n}$ 이지만, σ를 모르기 때문에 표본표준편차 s를 이용하여 추정하여야 한다. 표본분산 s^2은

$$s^2 = \frac{1}{40-1} \times \left\{ (2.6 - 1.715)^2 + \cdots + (1.2 - 1.715)^2 \right\} \doteq 0.2254$$

이므로, 표준오차의 추정치는 다음과 같다.

$$\frac{s}{\sqrt{n}} = \frac{\sqrt{0.2254}}{\sqrt{40}} \doteq 0.0751$$

6.3 모평균에 대한 구간추정

하나의 값으로 미지의 모수를 추정하는 점추정 대신에 모수의 참값을 포함할 것으로 기대되는 구간을 구하는 것이 더 바람직할 수 있다. 구간추정은 추정량의 확률분포를 이용하여 표본으로부터 모수의 참값을 포함하리라고 예상되는 구간을 구하여 제시하는 것이다. 가장 이상적인 구간추정은 표본을 추출하고 해당 표본을 이용하여 모수의 참값을 확실하게 포함하는 구간을 계산하는 것이다. 그러나 모수의 참값을 확실하게 포함하는 구간을 계산하는 것은 표본자료의 다양성 때문에 실제로는 불가능하다. 따라서 구간추정에서는 제안된 구간이 모수의 참값을 포함할 확률을 명시하게 되는데, 이 확률은 대개 90%, 95% 등을 사용하게 되고 이를 신뢰수준(level of confidence)이라 부른다. 더불어 이러한 신뢰수준 하에서 추정된 구간은 신뢰구간(confidence interval)이라 부른다.

신뢰구간의 개념을 이해하기 위하여 우선 모집단이 정규분포를 따르고 모분산 σ^2 이 알려져 있는 경우와 σ^2 이 알려져 있지 않은 보다 현실적인 경우로 나누어 모평균 μ 의 신뢰구간을 구하는 문제를 다루기로 하자. 우선, 5장에서 모집단이 정규분포를 따르면 크기 n 의 표본에 대한 표본평균 $\overline{X}$ 의 분포는 평균 μ 이고, 표준편차는 $\sigma/\sqrt{n}$ 인 정규분포가 됨을 배웠다. 따라서 신뢰수준 $1-\alpha$ 에서 표본평균 $\overline{X}$ 의 표준화변환식에 대한 확률식은 다음과 같이 표현 가능하다.

$$P\left(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha \tag{6.1}$$

여기서, $z_{\alpha/2}$ 는 표준정규분포의 상위 $\alpha/2$ 의 확률을 주는 값으로, 예를 들면 $\alpha = 0.05$ 인 경우 $z_{0.05/2} = z_{0.025} = 1.96$ 이 된다. 식 (6.1)의 괄호 안의 부등식을 μ 에 대해서 정리하면 다음과 같다.

$$P\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

따라서 모집단이 정규모집단이고 모분산 σ^2 을 알고 있을 경우, 다음의 구간을 모평균 μ 의 $100\times(1-\alpha)\%$ 신뢰구간이라 한다.

$$\left(\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ \ ,\ \ \overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

보기 6.2

평균이 μ 이고 표준편차가 $\sigma=8$ 인 정규분포를 따르는 모집단으로부터 크기 $n=25$ 인 확률표본을 추출하여 표본평균을 계산한 결과 $\overline{x}=42.7$ 이었다. 모평균 μ 에 대한 95% 신뢰구간을 구하라.

모집단의 확률분포가 정규분포이므로, 표본평균 $\overline{X}$ 의 확률분포 역시 정규분포이다. 더불어 모분산이 $\sigma^2=8^2$ 임을 알고 있다. 따라서 모평균 μ 의 95% 신뢰구간은

$$P\left(\overline{X}-z_{0.025}\frac{\sigma}{\sqrt{n}}<\mu<\overline{X}+z_{0.025}\frac{\sigma}{\sqrt{n}}\right)=0.95$$

와 같은 확률식으로 정의되며, 표본평균의 관측값 $\overline{x}=42.7$ 을 이용하면 모평균 μ 의 95% 신뢰구간은

$$\left(42.7-1.96\times\frac{8}{\sqrt{25}}\ \ ,\ \ 42.7+1.96\times\frac{8}{\sqrt{25}}\right)=(39.56,45.84)$$

가 된다.

지금까지 모집단의 분포가 정규분포이고 모분산 σ^2 을 아는 경우, 모평균 μ 의 신뢰구간에 대한 기본적 개념들을 설명하였다. 이젠 모분산 σ^2 을 모르는 보다 현실적인 상황에서의 신뢰구간을 생각해 보자. 이 경우에는 모분산 σ^2 의 추정량인 표본분산 S^2 을 사용하여야 한다. 따라서 표본평균 $\overline{X}$ 의 표준화변환식은

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \tag{6.2}$$

의 형태로 바뀌게 된다. 그런데 여기서 주의하여야 할 점은 설령 모집단이 정규분포를 따른다 하더라도 식 (6.2)의 변형된 표준화변환식의 확률변수의 확률분포는 표준정규분포와는 달라진다. 이런 경우에 그 변화된 확률분포를 t-분포라고 한다.

■ t-분포 (Student's t-distribution)

평균이 μ이고 분산이 σ^2인 정규모집단으로부터 추출된 크기 n의 표본을 $X_1, X_2, \cdots, X_n$이라 할 때, 이들에 대한 표본평균과 표본분산을 각각

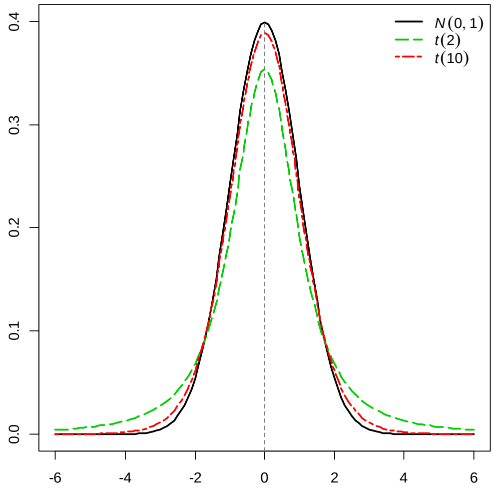
$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i,\quad S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

이라 정의하면, 확률변수

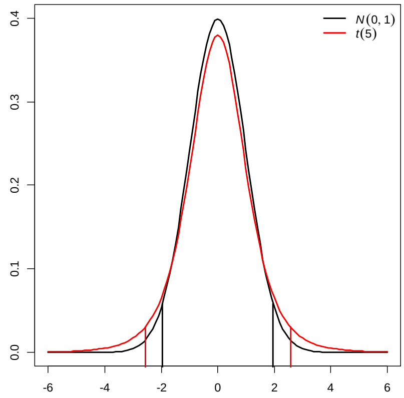
$$t=\frac{\overline{X}-\mu}{S/\sqrt{n}}$$

는 자유도(degree of freedom, df)가 $df=(n-1)$인 t-분포를 따른다 한다.

t-분포는 1908년에 영국의 고셋(W.S. Gosset)이라는 사람의 논문에서 소개되었는데, 이 논문은 스튜던트(Student)라는 필명으로 출간되었기에 그 필명을 따서 스튜던트의 t-분포라고 부르기도 한다. t-분포는 표준정규분포와 같이 0을 중심으로 좌우대칭인 종모양의 확률분포이다. 표준정규분포와 다른 점은 양쪽 꼬리부분에 상대적으로 많은 확률이 분포하여 표준정규분포보다 두꺼운 꼬리를 갖는다는 것이다.



[그림 6.1] t -분포와 표준정규분포



[그림 6.2] t -분포와 표준정규분포에서의 95% 확률의 구간

t -분포의 또 다른 특징은 자유도가 커지면 커질수록 점점 표준정규분포와 흡사해져간다는 것이다. [그림 6.1]에서 확인할 수 있듯이 자유도가 2인 t -분포보다 자유도가 10인 t -분포가 표준정규분포와 많이 유사한 모양을 가지게 된다. 이러한 성질을 반대로 생각하면, t -분포에서 자유도가 작으면 작을수록 신뢰수준 $100 \times (1-\alpha)$%에 따른 확률분포상의 구간의 길이는 길어지게 된다. [그림 6.2]에는 표준정규분포와 자유도가 5인 t -분포 상에서 95% 확률

의 구간이 나타나 있다. 그림 상에서 확인할 수 있듯이 자유도가 5인 t -분포 상에서의 구간의 길이가 정규분포 상에서의 구간의 길이보다 길다. 부록의 표를 이용하면 표준정규분포 상에서의 95% 확률 구간은 (-1.96, 1.96)이지만 자유도가 5인 t -분포 상에서는 (-2.571, 2.571)임을 확인할 수 있다. t -분포의 자유도는 $n-1$로 계산되므로 t -분포의 자유도가 작다는 것은 표본의 크기 n이 작아짐을 의미한다.

이러한 특징을 가지는 t -분포를 이용하여 신뢰수준 $1-\alpha$에서 표본평균 $\overline{X}$의 표준화변환식에 대한 확률식은 다음과 같이 표현 가능하다.

$$P\left(-t_{\alpha/2}(df) < \frac{\overline{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2}(df)\right) = 1-\alpha \tag{6.3}$$

여기서, $t_{\alpha/2}(df)$는 t -분포의 상위 $\alpha/2$의 확률을 주는 값이다. 식 (6.3)의 괄호 안의 부등식을 μ에 대해서 정리하면

$$P\left(\overline{X}-t_{\alpha/2}(df)\frac{S}{\sqrt{n}} < \mu < \overline{X}+t_{\alpha/2}(df)\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

와 같고, 모평균 μ의 $100\times(1-\alpha)\%$ 신뢰구간은 다음과 같다.

$$\left(\overline{X}-t_{\alpha/2}(df)\frac{S}{\sqrt{n}}\ ,\ \overline{X}+t_{\alpha/2}(df)\frac{S}{\sqrt{n}}\right) \tag{6.4}$$

우리는 앞서 [그림 6.1]을 통해 t -분포의 자유도가 커지면 표준정규분포와 유사해져간다는 것을 익힌 바 있다. 통상적으로 t -분포의 자유도가 30 이상이면 표준정규분포와 확률적 차이가 거의 없다고 말한다. 따라서 t -분포의 자유도가 30 이상이면 즉, 표본의 크기가 30보다 커지면 식 (6.4)의 신뢰구간은 t -분포를 이용하는 것과 표준정규분포를 이용하는 것의 차이가 거의 없기 때문에

$$\left(\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}\ ,\ \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

와 같이 계산할 수 있다.

■ 모평균 μ에 대한 $100\times(1-\alpha)\%$ 신뢰구간

정규모집단으로부터 추출한 크기 n의 표본에 대해

(i) 모분산 σ^2이 알려진 경우

$$\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ ,\ \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

(ii) 모분산 σ^2을 모르고 표본의 크기 n이 작은 경우($n \le 30$)

$$\left(\overline{X} - t_{\alpha/2}(df)\frac{S}{\sqrt{n}}\ ,\ \overline{X} + t_{\alpha/2}(df)\frac{S}{\sqrt{n}}\right)$$

(iii) 모분산 σ^2을 모르고 표본의 크기 n이 큰 경우($n > 30$)

$$\left(\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}\ ,\ \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

보기 6.3

[보기 6.1]의 1년생 붉은 소나무 묘목의 크기는 정규분포를 따른다고 가정하자. 전체 1년생 붉은 소나무 묘목의 평균 크기에 대한 90% 신뢰구간을 구하라.

우선 모집단은 정규분포를 따른다고 가정하였지만, 모분산 σ^2에 대한 정보는 주어져 있지 않다. 그러나 표본의 크기가 $n=40$이므로 충분히 크다. 따라서 모평균 μ에 대한 $100\times(1-\alpha)\%$ 신뢰구간은 다음과 같이 계산 가능하다.

$$\left(\overline{X}-z_{\alpha/2}\frac{S}{\sqrt{n}}\ ,\ \overline{X}+z_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

여기서 $1-\alpha=0.90$ 이므로 $z_{\alpha/2}=z_{0.05}=1.645$ 이다. 따라서 [보기 6.1]의 풀이로부터 표본평균과 표본표준편차

$$\overline{x}=1.715\,,\ s\doteq\sqrt{0.2254}=0.475$$

을 이용하면, 모평균 μ에 대한 90% 신뢰구간은 다음과 같다.

$$\left(1.715-1.645\times\frac{0.475}{\sqrt{40}}\ ,\ 1.715+1.645\times\frac{0.475}{\sqrt{40}}\right)=(1.591\,,1.839)$$

보기 6.4

식품의약품안전청에서는 어떤 생수의 단위량당 세균의 수치를 조사하고자 한다. 임의로 선택한 10개의 생수병을 검사한 결과 각 생수병에 대한 단위량당 세균의 수는 다음과 같았다.

175	190	215	198	184	207	210	193	196	180

각 생수병의 단위량당 세균의 수는 정규분포를 따른다고 가정했을 때, 해당 생수의 단위량당 평균 세균 수에 대한 95% 신뢰구간을 구하라.

우선 모집단은 정규분포를 따른다고 가정하였지만, 모분산 σ^2에 대한 정보는 주어져 있지 않다. 더불어 표본의 크기가 $n=10$ 이므로 충분히 크지 못하다. 따라서 모평균 μ에 대한 $100\times(1-\alpha)\%$ 신뢰구간은 다음과 같이 계산 가능하다.

$$\left(\overline{X}-t_{\alpha/2}(df)\frac{S}{\sqrt{n}}\ ,\ \overline{X}+t_{\alpha/2}(df)\frac{S}{\sqrt{n}}\right),\ df=10-1=9$$

현재 신뢰수준 $1-\alpha=0.95$ 이므로 $t_{\alpha/2}(9)=t_{0.025}(9)=2.262$ 이며, 표본에 대한 평균과 표준편차는 각각 $\overline{x}=194.8$ 과 $s\doteq13.14$ 이다. 따라서 모평균 μ에 대한 95% 신뢰구간은 다음과 같다.

$$\left(194.8-2.262\times\frac{13.14}{\sqrt{10}}\ ,\ 194.8+2.262\times\frac{13.14}{\sqrt{10}}\right)\doteq(185.4\,,204.2)$$

6.4 모비율에 대한 추정

6.2절에서 모평균 μ를 추정하고자 할 때, 직관적으로 가장 타당한 추정량은 표본평균 $\overline{X}$ 임을 언급한 바 있다. 이와 마찬가지로 모비율에 대해 가장 타당한 추정량으로는 표본비율을 사용할 수 있다. 예를 들어, 어느 제조회사에서 생산되는 제품에 대한 불량률을 알고자 한다면 실제로 측정해야 하는 모집단은 생산되는 제품 전체이다. 이것은 불가능하므로 생산되는 제품 중 임의로 n개의 제품을 추출하고 그 중 불량품의 개수를 파악하게 된다. 이러한 상황은 전체 불량률이 p인 모집단으로부터 상호독립적인 n개의 제품 중 불량품의 개수를 파악하는 것이므로, 추출한 n개의 제품 중 불량품의 개수를 확률변수 X라 하면 확률변수 X는 이항분포($\mathrm{Bin}\,(n, p)$)를 따르게 된다. 더불어 전체 불량률 즉, 모비율 p의 추정량인 표본비율은 다음과 같이 정의할 수 있다.

$$\hat{p} = \frac{X}{n}$$

따라서 확률변수 X가 이항분포를 따르므로 표본비율 $\hat{p}$의 기대값과 분산, 표준오차는 다음과 같다.

$$E(\hat{p}) = \frac{1}{n}E(X) = \frac{1}{n}np = p$$

$$Var(\hat{p}) = \frac{1}{n^2}Var(X) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

$$S.E.(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

이 경우, 표준오차에 있는 p는 추정하고자 하는 미지의 모수이기 때문에 이를 추정량인 $\hat{p}$로 대체하면 표준오차를 추정할 수 있다.

이제, n이 충분히 큰 경우 모비율 p의 구간추정을 생각해 보자. 이 경우 X는 근사적으로

평균이 np 이고 분산이 $np(1-p)$ 인 정규분포를 따른다. 따라서 n 이 충분히 큰 경우 표본비율 $\hat{p}$ 의 확률분포는 다음과 같이 정의할 수 있다.

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

그러므로 표본비율 $\hat{p}$ 의 표준화된 분포는

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$$

이 된다. 따라서 신뢰수준 $1-\alpha$ 에서 표본비율 $\hat{p}$ 의 표준화변환식에 대한 확률식은 다음과 같이 표현 가능하다.

$$P\left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}\right) = 1-\alpha \tag{6.5}$$

그러나 이때, 식 (6.5)의 괄호 안의 부등식을 p 에 대해서 단순하게 정리할 수 없으므로 분모 즉, 표본비율 $\hat{p}$ 의 표준오차에 대해 p 를 $\hat{p}$ 으로 대체하여 추정하여야 한다. 이와 같은 방법으로 식 (6.5)의 괄호 안의 부등식을 p 에 대해서 정리하면

$$P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < \mu < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1-\alpha$$

와 같고, 모비율 p 의 $100 \times (1-\alpha)\%$ 신뢰구간은 다음과 같이 구할 수 있다.

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \;,\; \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

■ 모비율 p에 대한 $100\times(1-\alpha)\%$ 신뢰구간

표본의 크기 n이 충분히 큰 경우, 모비율 p에 대한 $100\times(1-\alpha)\%$ 신뢰구간은 다음과 같다.

$$\left(\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\ \hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

보기 6.5

어느 공정과정에서 나오는 제품 중 100개를 임의로 추출하여 조사한 결과 6개가 불량품이었다. 이 공정과정에서 나오는 제품 전체의 불량률에 대한 95% 신뢰구간을 구하라.

표본의 크기 $n=100$은 충분히 크므로, 표본비율 $\hat{p}$은 근사적으로 정규분포를 따른다고 할 수 있다. 따라서 모비율 p에 대한 $100\times(1-\alpha)\%$ 신뢰구간은

$$\left(\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\ \hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

와 같이 계산할 수 있다. 신뢰수준은 $1-\alpha=0.95$이므로 $z_{0.025}=1.96$이며, 표본비율 $\hat{p}=6/100=0.06$이다. 따라서 모비율 p에 대한 95% 신뢰구간은 다음과 같다.

$$\left(0.06-1.96\times\sqrt{\frac{0.06\times0.94}{100}}\ ,\ 0.06+1.96\times\sqrt{\frac{0.06\times0.94}{100}}\right)\doteq(0.0135,0.1065)$$

6.5 R-프로그램 실습

[프로그램 6.1]은 [보기 6.4]의 세균 수에 대한 구간추정을 위한 R-프로그램이다. [보기 6.4]는 실제 데이터가 주어져 있으므로 R 내장함수 중의 하나인 t.test 함수를 이용할 수 있다. [결과 6.1]에서 95 percent confidence interval을 통해 95% 신뢰구간이 (185.4012, 204.1988)로 계산되었음을 확인할 수 있다. 만약, 90% 신뢰구간을 계산하고자 한다면 conf.level = 0.90으로 수정하면 된다.

[프로그램 6.1] [보기 6.4] 세균 수에 대한 구간추정

```
x = c(175,190,215,198,184,207,210,193,196,180)
t.test(x, conf.level = 0.95)
```

[결과 6.1] [프로그램 6.1]의 수행 결과

```
        One Sample t-test

data:  x
t = 46.8857, df = 9, p-value = 4.573e-12
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 185.4012 204.1988
sample estimates:
mean of x
    194.8
```

[프로그램 6.2]는 [보기 6.5]의 불량률에 대한 구간추정을 위한 R-함수이다. x는 성공 횟수 즉, 불량품의 개수를 의미하고, n은 전체 표본의 수, alpha는 신뢰수준 $1-\alpha$ 에서의 α 를 의미한다.

[프로그램 6.2] [보기 6.5] 불량률에 대한 구간추정

```
prob.CI = function(x, n, alpha){
          p.hat <- x/n
          error <- qnorm(1-alpha/2)*sqrt(p.hat*(1-p.hat)/n)
          res <- c(p.hat-error,p.hat+error)
          names(res) <- c("Lower","Upper")
          return(res)
       }
prob.CI(6,100,0.05)
```

[결과 6.2] [프로그램 6.2]의 수행 결과

```
     Lower      Upper
0.01345343 0.10654657
```

6.6 연습문제

6.1 코끼리의 체중에 대한 모표준편차가 $\sigma=15$파운드라고 알려져 있다. 임의로 선택된 코끼리 50마리의 체중을 측정한 결과 표본평균은 $\overline{x}=244$파운드이고, 표본표준편차는 $s=11$파운드이다. 모평균 μ에 대한 95% 신뢰구간을 구하라.

6.2 캘리포니아의 성인 506명을 대상으로 캘리포니아가 직면한 가장 큰 문제점을 조사한 결과, 그 중 400명이 '교육'이라고 답했다. 실제 교육이 가장 큰 문제라고 생각하는 캘리포니아의 성인의 비율에 관심이 있다.

(1) 모비율의 추정치와 표준오차를 구하라.

(2) 모비율 p에 대한 90% 신뢰구간을 구하라.

6.3 유럽 남성의 평균 체중에 대한 95% 신뢰구간을 구하기 위하여 스웨덴 남성 48명을 대상으로 조사한 결과, 표본평균은 71kg, 표본표준편차는 2.8kg이었다. 유럽 남성의 체중은 정규분포를 따른다고 가정하자.

(1) 모평균 μ에 대한 95% 신뢰구간을 구하라.

(2) 모분산이 9kg이라고 알려져 있다면, 모평균 μ에 대한 95% 신뢰구간은 어떻게 달라지는가?

6.4 한 제약회사에서 만드는 정신 안정제의 약효가 지속되는 시간은 정규분포를 따른다. 9명의 환자에게 이 약을 투여하고 약효가 지속되는 시간을 측정한 결과는 다음과 같다.

2.7, 2.8, 3.0, 2.3, 2.3, 2.2, 2.8, 2.1, 2.4

(1) 모평균의 추정치와 표준오차를 구하라.

(2) 모평균 μ에 대한 95% 신뢰구간을 구하라.

6.5 연습문제 6.4에서 표본평균과 표본표준편차는 동일하고, 표본의 크기가 49명으로 늘어난다면, 모평균 μ에 대한 95% 신뢰구간은 어떻게 달라지는가?

6.6 연습문제 6.4에서 모분산이 0.3시간으로 알려져 있다면, 모평균 μ에 대한 95% 신뢰구간은 어떻게 달라지는가?

6.7 마케팅 회사에서 주부 200명을 대상으로 조사한 결과, 그 중 120명이 물건을 구매할 때 제조회사를 가장 중요시 여긴다고 답했다. 실제 물건을 구매할 때 제조회사를 가장 중요시 여기는 주부의 비율에 관심이 있다. 모비율 p에 대한 95% 신뢰구간을 구하라.

6.8 동일 브랜드에서 판매되고 있는 여행용 가방 16개의 무게를 측정하였다. 여행용 가방의 무게는 정규분포를 따른다고 가정하자. 측정된 여행용 가방 무게의 표본평균은 2kg이고, 표본표준편차는 0.12kg이다. 모표준편차는 0.1kg으로 알려져 있다. 모평균 μ에 대한 90% 신뢰구간을 구하라.

6.9 부산대학교 학생 200명을 대상으로 조사한 결과에 따르면 일주일 동안 영어 공부를 하는 시간의 표본평균은 8.2시간, 표본표준편차는 2.2시간이었다. 모표준편차는 알려져 있지 않으며, 모집단은 정규분포를 따른다고 가정하자. 모평균 μ에 대한 90% 신뢰구간을 구하라.

6.10 두발 자전거를 배운 어린이 14명을 대상으로 두발 자전거를 배우기까지 걸린 기간을 조사하였다. 조사 결과, 표본평균은 6개월이고, 표본표준편차는 3개월이었다. 모집단은 정규분포를 따른다고 가정하자.

(1) 모평균 μ에 대한 90% 신뢰구간을 구하라.

(2) 모평균 μ에 대한 95% 신뢰구간을 구하라.

(3) 모평균 μ에 대한 99% 신뢰구간을 구하라.

7장

가설 검정 : 한 집단의 비교

R과 함께하는

통계학의 이해

7.1 귀무가설과 대립가설

어떤 모수에 관한 주장이나 추측 등을 표본자료에 의하여 증명하고자 할 때 통계적 가설 검정의 문제가 제기된다. 예를 들자면 [보기 6.1]에서 제기한 소나무의 품종 개발의 일환으로 성장이 좋다고 알려진 붉은 소나무에 관심이 있는 연구가 있다고 하자. 다른 소나무의 키는 평균 $\mu=1.9$m로 알려져 있다. 따라서 품종 개발자는 붉은 소나무의 키의 평균이 다른 소나무의 키의 평균 $\mu=1.9$m보다는 크다고 믿고 있다. 이를 증명하기 위하여 모든 1년생 붉은 소나무 묘목들로 이루어진 방대한 크기의 모집단으로부터 40그루의 표본자료 [표 6.1]을 임의 추출한 다음 표본평균 $\overline{X}$를 계산한다고 하자. 실제 모평균 μ가 1.9m보다 크다는 주장을 통계적으로 증명하기 위해서 이 결과를 어떻게 이용해야 하는지를 살펴보자.

제기된 주장의 타당성을 검정할 때, 통계학자들은 귀무가설(null hypothesis)과 대립가설(alternative hypothesis)이란 용어를 사용한다. 증명을 필요로 하는 주장을 대립가설(H_1)이라 하고 이 가설과 다른 주장 또는 기존의 주장을 귀무가설(H_0)이라 분류한다. 예를 들자면 붉은 소나무의 키의 모평균 μ가 1.9 m보다 크다고 제기된 주장은 대립가설 $H_1 : \mu > 1.9$이 된다. 귀무가설은 이 대립가설에 반대되는 두 가지 경우($\mu=1.9$과 $\mu<1.9$)에 대하여 세워진다. 그러나 연구자는 이 두 가지를 구분하지 않고 단순히 '붉은 소나무의 키 평균과 다른 소나무의 키의 평균에 있어서 차이가 나지 않는다'를 귀무가설로 하게 된다. 따라서 다음과 같은 가설을 세우게 된다.

$$H_0 : \mu=1.9, \quad H_1 : \mu>1.9$$

7.2 대표본의 모평균 검정

표본자료를 이용하여 모평균 μ가 1.9 m와 같다는 귀무가설 $H_0 : \mu = 1.9$를 기각할 것인가 말 것인가를 결정해야 한다. 이를 위한 기준으로 모평균 μ에 대응하는 임의로 추출된 40그루의 붉은 소나무 묘목의 표본자료로부터 계산된 표본평균 $\overline{X}$를 사용하는 것이 합리적으로 여겨진다. 그러면 '$\overline{X}$가 어떤 값일 때 H_0를 기각할 것인가'라는 의문을 가지게 될 것이다. 제기된 주장은 $\mu > 1.9$이므로 $\overline{X}$의 값이 클 때 H_0를 기각하고 제기된 주장 H_1이 옳다고 할 수 있을 것이다. $\overline{X}$의 값이 클 때를 일반화한 표현으로 $\overline{X} > c$를 쓸 수 있는데, 여기서 c는 귀무가설 H_0를 기각하는 기준값으로 이를 기각치(critical value)라 하며, $\overline{X} > c$의 구간은 기각역(rejection region 또는 critical region)이라 한다. 실제로 c는 표본 $X_1, X_2, \ldots, X_n$의 분포에 따라 결정된다.

여기서 기각치 c를 구하는 계산은 몇 단계를 거치는 번거로움이 있다. 일반적으로 표본 $X_1, X_2, \ldots, X_n$의 크기 n이 30보다 큰 대표본의 경우 단지 평균이 μ이고 표준편차가 σ인 모집단의 분포에 관계없이 $\overline{X}$의 분포는 근사적으로 평균이 μ이고 표준편차가 $\sigma/\sqrt{n}$인 정규분포를 따른다는 사실을 알고 있다. σ를 모를 경우 표본표준편차 s를 사용해도 된다. 따라서 $\overline{X}$의 표준화된 통계량

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \quad \text{또는} \quad Z = \frac{\overline{X} - \mu}{s/\sqrt{n}} \tag{7.1}$$

는 근사적으로 $N(0,1)$분포를 따르게 된다. 식 (7.1)의 표준화된 통계량 Z를 검정통계량(test statistic)이라 하며 z를 검정통계량의 관측값이라 하면, 이를 이용하여 기각역 $\overline{X} > c$ 대신에 $z > z_\alpha$를 정의할 수 있다. 여기서 기각치 z_α는 α의 확률을 정하면 부록의 표본정규분포표로부터 구할 수 있다. α는 유의수준(significance level)이라 하는데 귀무가설 H_0를 잘못 기각할 확률로 이 확률을 작게 하는 것이 가장 이상적이다. 대개 α를 0.10,

0.05, 또는 0.01과 같이 미리 정해진 수준으로 고정시키고 기각역을 결정하게 되는 것이다. 예를 들자면 $\alpha = 0.05$이면 $z_{0.05} = 1.645$가 된다.

■ 가설 검정과 관련된 용어들

귀무가설(null hypothesis, H_0) : 기존의 주장
대립가설(alternative hypothesis, H_1) : 증명을 필요로 하는 새로운 주장
기각치(critical value) : 귀무가설 H_0를 기각하는 기준값
기각역(critical region) : 기각치를 기준으로 귀무가설을 기각할 수 있는 범위
검정통계량(test statistic) : 가설 검정을 위한 모수의 점추정량
유의수준(significance level, α) : 귀무가설 H_0를 잘못 기각할 확률

보기 7.1

[보기 6.1]의 붉은 소나무 묘목의 키에 대한 표본자료 [표 6.1]에서 $n = 40$이고 표본평균 $\overline{x} = 1.715$와 표준편차 $s = 0.475$이므로 유의수준 0.05에서

$$H_0 : \mu = 1.9, \ \ H_1 : \mu > 1.9$$

를 검정하기 위한 기각역을 구하고 검정통계량의 관측값을 구한 후 기각역을 만족하는 지를 알아보라.

1) $\overline{X}$의 분포는 근사적으로 $N(1.9, \dfrac{0.475^2}{40})$이 되고 이를 표준화한 검정통계량은 근사적으로 표준정규분포를 따르며 다음과 같다.

$$Z = \frac{\overline{X} - 1.9}{0.475/\sqrt{40}} \simeq N(0,1)$$

2) 유의수준 $\alpha = 0.05$에서 기각치는 $z_{0.05} = 1.645$이므로 기각역은 다음과 같다.

$$z > 1.645$$

3) 표본평균이 $\overline{x} = 1.715$이므로 검정통계량의 관측값은

$$z = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{1.715 - 1.9}{0.475/\sqrt{40}} = -2.463$$

이다.

4) $z = -2.463$가 기각역을 만족하지 못하므로 귀무가설을 기각할 수 없게 된다.

5) 이 자료에 따르면 붉은 소나무 묘목의 키의 평균과 다른 소나무 묘목의 키의 평균과 같음을 알 수 있다.

7.3 단측 검정과 양측 검정

지금까지 [보기 7.1]에서 검정한 $H_0 : \mu = 1.9$, $H_1 : \mu > 1.9$에서 세워진 대립가설의 형태와는 반대로 $H_1 : \mu < 1.9$이 되면 기각역 $z < z_\alpha$를 사용하게 된다. 이들 두 가지 형태의 대립가설을 단측가설(one-sided hypothesis)이라 부르는데, 그 이유는 대립가설에서의 모수 μ의 값들이 귀무가설 $H_0 : \mu = 1.9$에서 주어진 값 $\mu_0(=1.9)$의 한 쪽에만 놓였기 때문이다. 따라서 이 대립가설들의 검정을 단측 검정(one-sided test 또는 one-tailed test)이라 부른다. 이와 대조적으로, 어떤 문제에서는 귀무가설 $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$을 검정하기도 한다. 이런 $H_1 : \mu \neq \mu_0$ 형태의 대립가설을 양측가설(two-sided hypothesis)이라 하며 이런 가설의 검정을 양측 검정(two-sided test)이라 한다. 이 경우 $\overline{X}$가 μ_0로부터 어느 쪽으로건 멀리 떨어진 값이면, 즉, Z가 너무 작거나 너무 크면 H_0를 기각한다. 유의수준 α인 검정을 하기 위하여 기각할 확률 α를 양쪽 꼬리에 균등하게 나누어 기각역

$$z \leq -z_{\alpha/2},\ z \geq z_{\alpha/2}$$

을 정한다. 이 기각역은 보다 간략한 기호로

$$|z| \geq z_{\alpha/2}$$

로 표현할 수 있다. 따라서 대립 가설이 단측이냐 양측이냐에 따라서 기각역도 단측 기각역이냐 양측 기각역이냐가 결정됨을 알 수 있다.

보기 7.2

닭 사료를 바꾸면 생산되는 계란의 콜레스테롤 양이 줄어든다는 연구가 있다. 이를 증명하기 위하여 표본자료 38개의 콜레스테롤 관측치로부터 계산된 표본평균과 표본표준편차가 각각 $261mg$과 $22mg$이라면 귀무가설과 대립가설을 세우고 유의수준 0.025에서 검정하라.

1) 계란의 콜레스테롤 양이 줄어드는가에 관심이 있으므로 귀무가설과 대립가설은 다음과 같이 세울 수 있다.

$$H_0 : \mu = 270,\ H_1 : \mu < 270$$

2) $n = 38$이고 관측값 $\overline{x} = 261,\ s = 22$를 이용하면 검정통계량의 관측값은

$$z = \frac{261 - 270}{22/\sqrt{38}} = -2.52$$

3) $\alpha = 0.025$이고 기각치는 $-z_{0.025} = -1.96$이므로 기각역은 대립가설의 형태를 따른 $z \le -1.96$이 된다.

4) 검정통계량의 관측값 $z = -2.52$는 기각역에 속하므로 유의수준 $\alpha = 0.025$에서 귀무가설은 기각된다.

5) 이 자료는 새로운 사료를 사용할 때 평균 콜레스테롤 함유량이 줄어든다는 양계학자의 주장을 강력히 뒷받침하고 있다.

보기 7.3

[보기 7.1]에서 40그루의 소나무 묘목의 키 자료에 대한 대립가설 $H_1 : \mu > 1.9$ 대신에 모평균 키가 $1.9m$와 다르다고 하는 대립가설 $H_1 : \mu \neq 1.9$를 유의수준 0.05에서 검정하여 보라.

1) 대립가설 $H_1 : \mu \neq 1.9$가 양측 검정이므로 유의수준을 $\alpha = 0.05$에서 기각치는 $z_{\alpha/2} = z_{0.025} = 1.96$이므로, 기각역은 다음과 같다.

$$|z| \ge 1.96$$

2) 표본평균 $\overline{x} = 1.715$와 표준편차 $s = 0.475$이므로 검정통계량의 관측값은

$$z = \frac{1.745 - 1.9}{0.475/\sqrt{40}} = -2.46$$

이다.

3) 양측 검정이므로 절대값 $|z| = 2.46$이 기각치 1.96보다 크기 때문에 기각역을 만족하므로 유의수준 $\alpha = 0.05$에서 H_0을 기각한다.

4) 양측 검정 결과 붉은 소나무 묘목과 다른 소나무 묘목의 키의 평균이 다름을 확인할 수 있다.

지금까지의 표본의 크기 $n > 30$으로 비교적 크다고 할 수 있는 대표본에서 모평균 μ의 가설 검정에 대해서 살펴보았다. 일반적으로 이런 검정을 대표본의 모평균에 대한 유의수준 α에서 $Z-$검정이라 하며 대립가설별로 요약하자면 다음과 같다.

■ 대표본의 모평균에 대한 유의수준 α에서 $Z-$검정

귀무가설	대립가설	기각역	검정의 형태
$H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$	$z \geqq z_\alpha$	단측 검정
	$H_1 : \mu < \mu_0$	$z \leqq -z_\alpha$	단측 검정
	$H_1 : \mu \neq \mu_0$	$\lvert z \rvert \geqq z_{\alpha/2}$	양측 검정

7.4 소표본의 모평균 검정

7.3절에서 식 (7.1)에 의한 귀무가설 $H_0 : \mu = \mu_0$에 대한 $Z-$검정은 표본의 크기가 큰 대표본에서 모집단의 분포에 무관하게 적용된 검정절차이다. 이와는 반대로 표본 $X_1, X_2, \ldots, X_n$의 크기가 작은 소표본($n \leqq 30$) 경우에도 모집단의 분포가 정규분포 $N(\mu, \sigma^2)$라는 가정이 있다면 검정통계량만 바꾸면 $Z-$검정과 전체적으로 동일한 절차로 검정을 할 수 있다. 즉, 소표본인 경우 귀무가설 $H_0 : \mu = \mu_0$에서 검정통계량

$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \tag{7.2}$$

는 자유도 $df = n-1$인 $t-$분포를 따른다. 따라서 실제 자료인 표본을 통한 식 (7.2)의 검정통계량의 관측값 t에 의한 모평균에 대한 가설 검정은 다음과 같이 정리되며 이를 소표본의 모평균에 대한 유의수준 α에서 $t-$검정이라 한다.

■ 소표본의 모평균에 대한 유의수준 α에서 $t-$검정

귀무가설	대립가설	기각역	검정의 형태
$H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$	$t \geqq t_\alpha$	단측 검정
	$H_1 : \mu < \mu_0$	$t \leqq -t_\alpha$	단측 검정
	$H_1 : \mu \neq \mu_0$	$\lvert t \rvert \geqq t_{\alpha/2}$	양측 검정

보기 7.4

어느 도시 위생국에서는 어떤 생수의 단위량당 세균의 평균 숫자가 안전수준인 200 이내인지를 조사하고자 한다. 한 조사원이 10개의 표본자료를 검사한 결과 세균의 수는 다음과 같고 정규분포를 따른다.

175, 190, 215, 198, 184, 207, 210, 193, 196, 180

귀무가설과 대립가설을 세우고 이를 유의수준 $\alpha = 0.01$에서 검정하여 보라.

1) μ를 이 생수 단위량당 세균의 평균 숫자라고 하자. 이 생수가 안전수준이라면 $\mu < 200$이고 조사원은 이 가설을 뒷받침하는 강력한 증거를 찾고자 한다. 그러므로 귀무가설과 대립가설은 다음과 같다.

$$H_0 : \mu = 200, \quad H_1 : \mu < 200$$

2) 표본의 크기가 $n = 10$인 소표본이며 정규분포를 따르므로 t-검정을 실시할 수 있다. 표본자료로부터 표본평균, 표준편차, 검정통계량의 관측값을 계산하면 다음과 같다.

$$\bar{x} = 194.8, \quad s = 13.14$$

$$t = \frac{194.8 - 200}{13.14/\sqrt{10}} = \frac{-5.2}{4.156} = -1.25$$

3) 대립가설이 단측인 $H_1 : \mu < 200$의 형태에 따른 유의수준 $\alpha = 0.01$에서 자유도 $df = n - 1 = 9$인 기각치는 $-t_{0.01} = -2.821$이고 기각역은 $t \le -2.821$이 된다.

4) 검정통계량의 관측치 $t = -1.25$는 -2.821보다 크므로 기각역 $t \le -2.821$을 만족하지 않는다. 이는 유의수준 $\alpha = 0.01$에서 귀무가설 $H_0 : \mu = 200$이 기각되지 않음을 말한다.

5) 10개의 표본자료는 모평균이 안전수준 이내에 있다는 증거가 되지 못한다.

7.5 모비율의 검정

새로 개발된 공법의 불량률과 부산시 노동력 인구 중 실업자 비율에 대한 검정은 모비율 p에 대한 가설 검정에 해당한다. 소표본의 경우는 개념적으로 앞서 정리해온 모평균의 가설 검정의 절차와 달리 복잡하므로 생략하고 대표본인 경우만 다루기로 한다.

모비율 p인 모집단으로부터 모비율과 관련된 특성의 경우의 수를 X라 하면 귀무가설 $H_0 : p = p_0$에서 모비율의 추정량 $\hat{p}$은 n이 클 때 근사적으로 다음과 같이 정규분포를 따른다.

$$\hat{p} = \frac{X}{n} \simeq N(p_0, \sqrt{p_0 q_0 / n}) \tag{7.3}$$

따라서 식 (7.3)으로부터 표준화된 검정통계량은 식 (7.4)의 근사적으로 표준정규분포를 따른다.

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} \approx N(0,1) \tag{7.4}$$

표본을 통한 대표본의 모비율 p에 대한 식 (7.4)의 검정통계량의 관측값 z에 의한 유의수준 α에서 $Z-$검정은 모평균 μ에 대한 $Z-$검정과 유사한 방법으로 다음과 같다.

■ 대표본의 모비율에 대한 유의수준 α에서 $Z-$검정

귀무가설	대립가설	기각역	검정의 형태
$H_0 : \mu = \mu_0$	$H_1 : p > p_0$	$z \geqq z_\alpha$	단측 검정
	$H_1 : p < p_0$	$z \leqq -z_\alpha$	단측 검정
	$H_1 : p \neq p_0$	$\lvert z \rvert \geqq z_{\alpha/2}$	양측 검정

보기 7.5

5년 전의 인구조사에 따르면 어떤 도시의 세대 중 20%가 빈곤층으로 밝혀졌다. 이 비율이 변하였는지를 보기 위하여, 400세대를 임의 추출하여 조사한 결과 70세대가 빈곤층이었다. 이 표본 자료에 의하면 현재 이 도시의 빈곤층 비율이 5년 전과 달라졌다고 할 수 있는지를 살펴보라.

1) 모비율 p를 현재 이 도시의 빈곤층 비율이라 하면 p가 5년 전 빈곤층 비율 0.20과 다른지에 대한 증거를 찾고 있으므로, 가설은 다음과 같다.

$$H_0 : p = 0.20,\ H_1 : p \neq 0.20.$$

2) 표본의 크기 $n = 400$은 충분히 크므로 대표본에 해당하며 모비율 p의 추정치는 $\hat{p} = 70/400$이고 $H_0 : p = 0.20$에서 검정통계량의 관측값은 다음과 같다.

$$z = \frac{(70/400) - 0.2}{\sqrt{0.2 \times 0.8/400}} = \frac{0.175 - 0.2}{0.020} = -1.25$$

3) 대립가설이 양측인 $H_1 : p \neq 0.20$이므로 유의수준 $\alpha = 0.05$에서 기각치는 $z_{0.025} = 1.96$이고 기각역은 $|z| \geq 1.96$이다.

4) 검정통계량 관측값 $z = -1.25$의 절대값 $|z| = 1.25$은 기각역을 만족하지 못하므로 귀무가설 $H_0 : p = 0.20$를 기각하지 못한다.

5) 검정의 결과 현재 빈곤층의 비율이 5년 전과 달라졌다고는 할 수 없다.

7.6 오류와 유의확률

지금까지 살펴본 통계적 가설 검정의 결과는 유의수준 α에서 모평균 μ와 모비율 p에 관한 귀무가설 H_0를 기각하거나 기각하지 않거나 둘 중의 한 가지였다. 이럴 경우 특히 모평균 μ나 모비율 p를 직접 관측할 수 없어 표본을 통해 추정된 추정량을 통하여 귀무가설과 대립가설의 옳고 그름을 판단해야 하므로 오류를 범할 수 있다. [표 7.1]에는 가설 검정에서 실제 사실과 이를 검정하기 위한 표본에 의해 결정한 것에 따라 발생하는 네 가지 경우를 요약하고 있다.

[표 7.1] 가설 검정의 결과

실제 사실	표본에 의한 결정	
	H_0 기각	H_0 채택
H_0 참	잘못 결정 (제1종 오류)	옳은 결정
H_1 참	옳은 결정	잘못 결정 (제2종 오류)

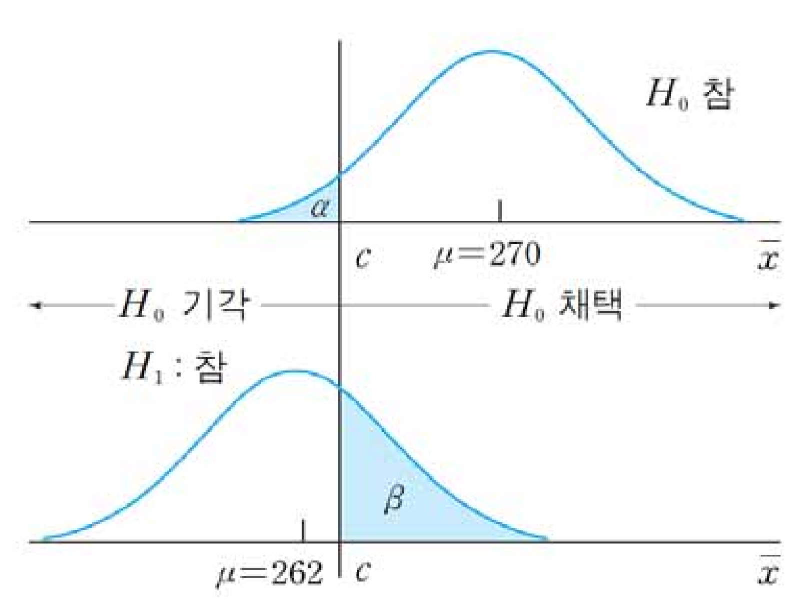
[표 7.1]에서와 같이 만약에 실제 사실로 H_0가 참임에도 표본자료에 의한 검정결과 H_0를 기각한다면 그것이 잘못된 결정이며 제1종 오류(type I error)를 범하게 된다. 반대로 H_1이 참인 경우 H_0를 채택하게 되면 제2종 오류(type II error)를 범하게 된다. 특히 제1종 오류를 범할 확률을 유의수준이라 하고 α로 나타내고 제2종 오류를 범할 확률을 β로 나타낸다.

7.2절에서 귀무가설 H_0를 기각하는 기준값이 되는 기각치 c에 대하여 설명한 바 있다. 특히, [보기 7.2]의 계란의 콜레스테롤 함유량 문제의 경우 모평균 μ에 대한 귀무가설 $H_0 : \mu = 270$에 대하여 대립가설이 단측인 $H_1 : \mu < 270$이므로 기각역의 형태는 $\overline{X} \leq c$이었다. 이 기각역은 귀무가설이 옳다는 가정 아래 이를 기각하는 구간이며 그 반대는 $\overline{X} > c$가 된다. 따라서 제1종 오류를 범할 확률인 유의수준 α와 제2종 오류를 범할 확률을

β를 각각 구해보면 다음과 같다.

$$\alpha = P[\overline{X} \leq c],\ \ \beta = P[\overline{X} > c]$$

물론, 확률 β는 $H_1 : \mu < 270$이 참인 범위의 실제 μ값에 따라서 달라진다. [그림 7.1]의 첫 번째 그림에서 음영 부분은 기각치 c를 기준으로 $H_0 : \mu = 270$가 참인 $\mu = 270$인 경우 제1종 오류를 범할 확률 α를 나타내고, 두 번째 그림에서 음영 부분은 $H_1 : \mu < 270$이 참인 경우 중 $\mu = 262$일 때 제2종 오류를 범할 확률 β를 나타낸다.



[그림 7.1] 제1종 오류와 제2종 오류의 확률 α와 β

[그림 7.1]에 의하면 두 가지 종류의 오류를 범할 확률 α와 β가 동시에 최소가 되도록 기각치 c를 결정하는 것은 명백히 불가능하다. c값을 왼쪽으로 움직이면 α는 작아지지만 β가 커지고, c값을 오른쪽으로 움직이면 그 반대의 현상이 일어난다. 이러한 문제점과 더불어 가설 검정에서 기존의 사실이 옳음에도 이를 기각하는 제1종 오류가 제2종 오류보다 더 심각한 오류라는 사실 때문에 α를 0.10, 0.05, 또는 0.01과 같이 미리 정해진 수준으로 고정시키고

기각역을 결정하게 되는 것이다.

[보기 7.2]의 새로운 사료에 의한 계란의 콜레스테롤 양에 관한 귀무가설과 대립가설 $H_0 : \mu = 270,\ H_1 : \mu < 270$의 $Z-$검정은 유의수준 $\alpha = 0.025$에서 행해졌다. 기각치가 $-z_{0.025} =-1.96$으로 검정통계량의 관측값 $z =-2.52$가 기각역 $z \le -1.96$에 속하였기 때문에 H_0를 기각하였다. 그리고 유의수준 $\alpha = 0.025$보다 큰 보편적인 유의수준 $0.05, 0.1$을 사용하면 당연히 귀무가설을 기각한다. 이번에는 $\alpha = 0.025$보다 작은 유의수준 $\alpha = 0.01$을 선택하더라도 기각치가 $-z_{0.01} =-2.33$과 기각역 $z \le -2.33$으로 인해 같은 검정 결과를 가져온다. 그러나 이번에는 훨씬 작은 유의수준 $\alpha = 0.005$에서는 기각치가 $-z_{0.005} =-2.57$로 검정통계량의 관측값 $z =-2.52$가 기각역 $z \le -2.57$을 벗어나게 되어 귀무가설을 기각하지 못하게 된다. 이와 같이 유의수준에 따라 검정의 결과가 달라지며 여러 상황에서 적절한 유의수준을 정하기는 것이 쉽지 않다.

기왕이면 검정통계량의 관측값 $z =-2.52$를 기각치로 하고 귀무가설을 기각할 확률

$$P[Z \le -2.52] = 0.0059$$

을 이용하는 것도 좋은 방법이 되며 실제 이 확률은 검정통계량의 관측값 $z =-2.52$를 이용하여 H_0을 기각할 수 있는 최소 유의수준이 된다. 이를 관측값 z의 유의확률(significance probability) 또는 p-값(p-value)이라 부른다. 아주 작은 p-값=0.0059는 H_0를 기각하는 강력한 기준이 되고 따라서 검정 결과는 통계적으로 아주 유의하게 된다. 일반적으로 p-값은 H_0가 참일 때 검정통계량이 실제 표본을 통한 검정통계량의 관측값과 같거나 더 지나친 값을 취할 확률이다.

가설 검정에서 미리 정한 유의수준 α와 관련된 검정통계량의 분포와 관측치를 활용한 기각치와 기각역을 활용하는 것 대신 검정통계량의 관측값의 유의확률인 p-값을 구하고 가설 검정의 상황에 따라 정한 유의수준 α보다 작으면 귀무가설 H_0를 기각하게 되는 검정 방법을 고려할 수 있다.

따라서 대표본의 모평균과 모비율의 검정에 대해선 검정통계량의 관측값 z에 의해 대립가설의 형태에 따라 각각 구한 유의확률 $P(Z \geq z)$, $P(Z \leq z)$, $P(|Z| \geq z)=2P(Z \geq z)=2P(Z \leq -z)$인 p-값에 의해서도 검정을 하게 된다. 소표본의 모평균의 검정에서는 검정통계량의 관측값 t에 의해서 p-값으로 대립가설의 형태에 따라 구한 $P(T \geq t)$, $P(T \leq t)$, $P(|T| \geq t)=2P(T \geq t)=2P(T \leq -t)$을 고려하게 된다.

■ p-값(p-value)의 정의와 활용

- H_0가 참일 때 검정통계량이 실제 표본을 통한 검정통계량의 관측값과 같거나 더 지나친 값(크거나 작은 값)을 취할 확률이다.
- p-값 $< \alpha \Rightarrow$ 귀무가설을 기각한다.

그러나 검정통계량의 관측값 분포표에서 정확하게 주어져 있지 않아 유의확률 p-값을 근사적으로 구할 수는 있지만 정확하게 계산되지는 못한다. 그러므로 7장을 포함하여 8장 등 본 저서의 본문 중 여러 보기에서는 기각치와 기각역을 이용한 검정 방법을 다루며 다만 7.8절과 같이 각 장의 마지막 절인 [R-프로그램 실습]에서는 수행 결과 p-값을 제공하므로 이를 활용한 검정을 보충하여 설명하기로 하자.

[보기 7.6]에서는 [보기 7.3]과 [보기 7.5]를 이용하여 유의확률 p-값의 계산에 대한 예와 검정방법을 보여주고 있다

보기 7.6

[보기 7.3]의 붉은 소나무 묘목의 키에 대한 표본자료 $H_0 : \mu = 1.9$, $H_1 : \mu \neq 1.9$에 대한 검정과 [보기 7.5]의 현재 이 도시의 빈곤층 모비율 p에 대한 p가 5년 전 빈곤층 비율 0.20과 다른지에 대한 가설 $H_0 : p = 0.20$, $H_1 : p \neq 0.20$에 대한 검정을 유의수준 $\alpha = 0.05$에서 기각역을 이용한 검정을 하고 있다. 이번에는 p-값을 계산하여 그 결과를 해석하라.

1) 40그루 소나무 묘목의 키의 자료에서 검정통계량의 관측치가 $z=-2.46$이고 양측 대립가설 $H_1: \mu \neq 1.9$이므로 유의확률(p-값)은

$$P[|Z| \geq 2.46] = P[Z \leq -2.46] + P[Z \geq 2.46]$$
$$= 2 \times P[Z \geq 2.46]$$
$$= 2 \times 0.0069 = 0.0138$$

이다. 이 p-값 = 0.0138이 $\alpha = 0.05$보다 작아서 H_0가 기각된다.

2) [보기 7.5]의 경우 검정통계량의 관측치가 $z=-1.25$이고 양측 대립가설 $H_1: p \neq 0.20$이므로 유의확률(p-값)은

$$P[|Z| \geq 1.25] = P[Z \leq 1.25] + P[Z \geq 1.25]$$
$$= 2 \times P[Z \geq 1.25]$$
$$= 2 \times 0.1056 = 0.2112$$

이다. p-값 = $0.2112 > \alpha = 0.05$이므로 H_0를 기각할 수 있는 증거는 매우 약하다.

7.7 R-프로그램 실습

[프로그램 7.1]은 표본의 크기가 큰 대표본의 경우인 [보기 7.1]의 붉은 소나무 묘목의 키의 모평균에 대한 가설 $H_0 : \mu = 1.9$, $H_1 : \mu > 1.9$ 검정을 위한 z.test( ) 함수를 활용한 R-프로그램이다. 이 함수를 사용하기 위해서는 프로그램에서 자료 x=c(2.6, 1.9, …, 1.2), 모평균 1.9, 표본분산 $s^2 = 0.475^2 = 0.2256$이 필요하다. 실제 프로그램에서 <z=z.test(x, 1.9, 0.2256)>에 해당한다. R-프로그램에서는 <pvalue=1-pnorm(z)>를 통하여 대립가설 $H_1 : \mu > 1.9$에 대해 기각치를 활용한 검정을 제공하는 대신에 $p-$값을 제공할 수 있다.

[결과 7.1]의 <-2.463384>는 검정통계량의 관측값 z에 해당하며 <0.9931184>가 $p-$값으로 확률값 $P[Z \geq -2.46] = 1 - P[Z > 2.46]$에 해당한다. 유의수준 $\alpha = 0.05$보다도 훨씬 크므로 귀무가설 $H_0 : \mu = 1.9$을 기각하지 못한다.

[프로그램 7.1] [보기 7.1] 대표본의 모평균에 대한 $Z-$검정

```
z.test = function(x, mu, var){
zeta = (mean(x) - mu)/(sqrt(var/length(x)))
  return(zeta)}
x=c(2.6, 1.9, 1.8, 1.6, 1.4, 2.2, 1.2, 1.6,
1.6, 1.5, 1.4, 1.6, 2.3, 1.5, 1.1, 1.6,
2.0, 1.5, 1.7, 1.5, 1.6, 2.1, 2.8, 1.0,
1.2, 1.2, 1.8, 1.7, 0.8, 1.5, 2.0, 2.2,
1.5, 1.6, 2.2, 2.1, 3.1, 1.7, 1.7, 1.2)
z=z.test(x, 1.9, 0.2256)
z
pvalue=1-pnorm(z)
pvalue
```

[결과 7.1] [프로그램 7.1]의 수행 결과

```
-2.463384
 0.9931184
```

[프로그램 7.2]는 소표본의 모평균에 대한 $t-$검정을 위한 t.test( ) 함수를 활용한 R-프로그램이다. 실제 이 프로그램은 [보기 7.4]에서 생수 단위량당 세균의 평균 숫자를 모평균이라 하고 이에 대한 $H_0 : \mu = 200,\quad H_1 : \mu < 200$의 검정을 보여주고 있다. 특히, t.test( ) 함수에서 지정하는 옵션 <x>는 입력자료 <x=c(175, 190, …, 180)>이고 <mu=200>은 귀무가설의 모평균 그리고 <alternative="less">는 단측 대립가설 $H_1 : \mu < 200$에 따른 옵션이다. 참고로 또 다른 단측 대립가설 $H_1 : \mu > 200$이면 <alternative = "greater">를 지정하고 양측 대립가설 $H_1 : \mu \neq 200$이면 <alternative = "two.sided">를 지정하면 된다. 이 옵션은 R-프로그램에서 제공하는 모든 가설 검정에서 대립가설의 형태에 따라 동일하게 사용된다.

결과에 대한 해석은 [보기 7.4]와 [보기 7.6]의 $p-$값 계산을 참고하면 되나 [결과 7.2]의 <t = -1.2516>는 검정통계량의 관측값이고 자유도가 <df = 9>임을 나타내고 있다. 특히, <p-value = 0.1211>는 $p-$값으로 유의수준 $\alpha = 0.05$보다도 큰 경우를 나타내며 귀무가설 $H_0 : \mu = 200$을 기각하지 못함을 보여준다.

[프로그램 7.2] [보기 7.4] 소표본의 모평균에 대한 $t-$검정

```
x=c(175, 190, 215, 198, 184, 207, 210, 193, 196, 180)
t.test(x, mu=200, alternative="less")
```

[결과 7.2] [프로그램 7.2]의 수행 결과

```
          One Sample t-test

data:  x
t = -1.2516, df = 9, p-value = 0.1211
alternative hypothesis: true mean is less than 200
95 percent confidence interval:
    -Inf 202.4162
sample estimates:
mean of x
   194.8
```

[프로그램 7.3]은 [보기 7.5]에서 도시의 빈곤층 모비율이 5년 비율 0.20과 다른지에 대한 가설 $H_0 : p = 0.20,\ H_1 : p \neq 0.20$에 대한 prop.test( ) 함수를 활용한 Z-검정을 보여주고

있다. 특히, prop.test( ) 함수에서 지정하는 옵션으로 <70>은 조사 결과 빈곤 세대수, <400>은 총 조사 세대수, <p=0.2>는 귀무가설 $H_0 : p = 0.20$의 모비율 값이다. <alternative="two.sided">는 양측 대립가설 $H_1 : p \neq 0.20$에 대한 옵션이며 이미 [프로그램 7.2]에서 설명한바 있다. 끝으로 옵션 <correct=FALSE>는 Yates'의 연속 보정 없이 식 (7.4)의 근사적으로 표준정규분포를 따르는 검정통계량 Z 의 제곱 통계량 $X^2 = Z^2$ 은 자유도 1인 X^2-분포가 된다는 수리적인 사실을 이용하며 실제로 [보기 7.5]에서 검정통계량의 관측값의 제곱인 $z^2 = (-1.25)^2$이 [결과 7.3]에서 <X-squared = 1.5625>에 해당한다. <p-value = 0.2113>은 $\alpha = 0.05$보다 크므로 H_0를 기각할 수 있는 증거는 매우 약함을 보여주고 있다.

[프로그램 7.3] [보기 7.5] 대표본의 모비율에 대한 검정

```
prop.test(70, 400, p=0.2, alternative="two.sided",correct=FALSE)
```

[결과 7.3] [프로그램 7.3]의 수행 결과

```
             1-sample proportions test without continuity correction

data: 70 out of 400, null probability 0.2
X-squared = 1.5625, df = 1, p-value = 0.2113
alternative hypothesis: true p is not equal to 0.2
95 percent confidence interval:
 0.1409042 0.2152788
sample estimates:
    p
0.175
```

참고로 Yates'의 연속 보정은 카이제곱 검정통계량을 계산할 때 다음의 공식

$$X^2 = Z^2 = \left[\frac{|\hat{p} - p_0| - \frac{1}{2n}}{\sqrt{p_0 q_0 / n}} \right]^2$$

을 이용하게 하며 prop.test( ) 함수에서 <correct=TRUE>를 지정하면 된다.

7.8 연습문제

7.1 다음과 같은 여러 가지의 주장이 제시되었다. 각 경우에 귀무가설(H_0)과 대립가설(H_1)을 나타내고, 검정의 형태가 단측인지 양측인지 말하라.

(1) 은퇴 전 일하는 평균 기간은 34년과 차이가 있다.

(2) 고등학교 3학년 학생 중 음주를 하는 학생은 29%와 차이가 있다.

(3) 대학교 졸업자의 평균 초기 연봉은 아무리 많아도 3,000만 원에 못 미친다.

(4) 여성들의 유방암 발생 가능성은 11% 이상이다.

(5) 대통령 선거의 투표율이 아무리 높아도 80%를 넘지 못한다.

(6) 유럽인들의 올해 휴가 기간은 작년 평균 휴가 기간인 6주와 차이가 있다.

(7) 사립 대학교의 평균 등록금은 735만 원보다 많다.

(8) 도시에서 떨어져 사는 것을 선호하는 미국인은 절반이 아니다.

7.2 연습문제 7.1의 (1)~(4)에서 제1종 오류와 제2종 오류를 범하는 상황은 어떠한 경우인지 말하라.

7.3 부산대학교의 통계학과 학생들은 평균적으로 7시간보다 적게 잔다고 알려져 있다. 이를 확인하기 위하여 실제 부산대학교 통계학과 학생들 중 22명의 조사하였더니 평균은 6.24시간, 표준편차는 1.93시간으로 나타났다.

(1) 귀무가설 H_0와 대립가설 H_1을 서술하라.

(2) 이 검정은 어떤 분포를 사용하는가?

(3) 검정통계량을 구하여라.

(4) 유의수준 5%에서 어떠한 결론을 내릴 수 있는가?

(5) 실제 귀무가설이 참일 때, (4)의 결정은 옳은 결정인가? 만일 잘못된 결정이라면 어떠한 종류의 오류를 범했는가?

7.4 한 신문기사에 따르면, 중국에서의 남녀 출생 비율은 100:114(여자=46.7%)라고 한다. 그러나 이 비율이 잘못되었다고 주장하기 위하여 중국에서 최근 출생한 아기 150명을 조사한 결과, 남자가 90명이고 여자가 60명이었다.

(1) 귀무가설 H_0와 대립가설 H_1을 서술하라.

(2) 검정통계량을 구하여라.

(3) 유의수준 5%에서 어떠한 결론을 내릴 수 있는가?

7.5 공인 간호사의 평균 연봉은 69,110달러라고 알려져 있다. 캘리포니아의 공인 간호사들의 평균 연봉은 69,110달러보다 높을 것이라고 생각하고, 이를 확인하기 위하여 캘리포니아의 공인 간호사 41명을 조사하였다. 조사 결과, 평균은 71,121달러, 표준편차는 7,489달러로 나타났다. 유의수준 5%에서 검정하여라.

7.6 연습문제 7.5의 검정은 유의수준 $\alpha = 0.05$에서 기각역을 이용한 검정을 하고 있다. 이번에는 p-값을 계산하여 그 결과를 해석하라.

7.7 부산대학교 교내 신문에 따르면 부산대학교 학사 학위를 얻는데 평균적으로 4.5년이 걸린다고 한다. 이를 확인하기 위하여 49명의 학생을 조사하였더니 평균이 5.1년, 표준편차가 1.2년이었다. 실제로 학사 학위를 얻는데 걸리는 기간이 4.5년과 차이가 있다고 할 수 있는지 유의수준 1%에서 검정하여라.

7.8 '성인 남성의 평균 IQ는 107이고, 브라운송어의 평균 IQ는 4이다. 그런데 왜 그들은 브라운송어를 잡지 못하는가?'라는 신문 광고 문구가 있다. 이 문구를 본 한 물고기 심리학자는 브라운송어의 평균 IQ가 4보다 높을 것으로 예상하여, 브라운송어 12마리를 잡아서 실험하였다. 측정된 브라운송어의 IQ는 다음과 같다.

5, 4, 7, 3, 6, 4, 5, 3, 6, 3, 8, 5

(1) 유의수준 5%에서 기각역을 이용하여 검정하여라.

(2) p-값을 계산하여 유의수준 5%와 1%에서 각각 검정하여라.

7.9 연습문제 7.8의 검정에서 대립가설 H_1: $\mu > 4$ 대신에 브라운송어의 평균 IQ가 4와 다르다고 하는 대립가설 H_1: $\mu \neq 4$를 검정하여라.

(1) 유의수준 5%에서 기각역을 이용하여 검정하여라.

(2) 유의수준 5%에서 p-값을 이용하여 검정하여라.

7.10 한 조사에 따르면 작년 청소년의 흡연율은 14%였다고 한다. 올해 청소년의 흡연율을 확인하기 위하여 조사한 결과, 70명 중 9명이 흡연하는 것으로 나타났다. 청소년의 흡연율이 14%로 유지되고 있는지, 아니면 감소하였는지를 유의수준 5%에서 검정하여라.

8장

독립표본과 대응표본 : 두 집단의 비교

R과 함께하는

통계학의 이해

8.1 두 집단의 비교

통계적으로 두 집단 간의 차이를 이해하기 위해 비교연구(comparative study)를 하게 된다. 예를 들어 연구자들이 체중 감소에 효과가 있는 새로운 운동처방 프로그램에 관심이 있다고 하자. 이 경우 기존의 운동처방 프로그램을 적용한 그룹과 새로운 운동처방 프로그램을 적용한 두 그룹에 대한 체중 감소를 비교하게 된다. 또 다른 예로 아스피린이 심장발작을 예방한다는 효과에 관심이 있다고 하자. 지난 몇 년간, 신문이나 잡지에 두 그룹에 대한 다양한 아스피린 연구가 보고되었다. 이 경우 일반적으로 아스피린을 복용한 그룹과 위약(placebo)을 복용한 그룹으로 나누어 여러 해 동안 두 그룹의 심장발작 비율을 비교하게 된다. 추가적으로 비교연구로 다양한 다이어트와 운동 프로그램을 서로 비교하거나, 정치인들에게 투표하는 사람들의 고소득 및 저소득 그룹 빈도수에 따른 비율에 관심을 가지는 경우이다. 두 그룹을 비교하기 위해서 필요한 자료의 수집 방법으로 1장에서 언급한 바 있지만 실험계획(experiment design)과 표본조사나 여론조사에 의해서 이루어진다. 운동처방 프로그램이나 심장발작 연구와 같은 자료는 철저한 실험계획에 의해서 얻어지며 정치인의 유권자에 대한 소득이나 지지도에 대한 자료는 여론조사를 통해 얻어진다.

어느 방법이든지 두 그룹은 서로가 관련이 없는 두 모집단으로부터 얻어진 독립표본(independent sample)이거나 아니면 한 모집단으로부터 얻어진 대상에 대하여 두 번 반복 측정하여 얻어진 두 표본인 대응표본(matched pairs sample)이 된다. 예를 들자면 기존의 운동처방 프로그램을 적용한 그룹과 새로운 운동처방 프로그램을 적용한 그룹이나 아스피린을 복용한 그룹과 위약을 복용한 그룹으로 나누어 여러 해 동안 두 그룹의 심장발작 비율을 조사하는 경우의 두 그룹은 서로에게 영향을 끼칠 수 없는 독립표본이 된다. 물론 실험하기 전 두 그룹의 신체적 조건(혈압, 몸무게)은 유사하도록 확보되어야 함은 당연하다. 그러나 실험에 따라 두 그룹의 유사한 신체적 조건을 보장하기 어려울 경우 대안으로 동일한 대상에 대하여 시간을 두고 측정하는 대응표본을 얻게 된다. 이 경우 먼저 실시하는 운동처방 프로그램이나 복용하는 약의 효과가 남아 있지 않도록 주의해야 한다.

■ 독립표본과 대응표본

독립표본 : 서로가 관련이 없는 두 모집단으로부터 각각 얻어진 두 표본
대응표본 : 한 모집단으로부터 얻어진 동일한 각 대상에 대하여 두 번 반복 측정하여 얻어진 두 표본

앞서 살펴본 바와 같이 실험계획에 의하여 비교하고자 하는 두 그룹을 사전에 나누거나 동일한 대상에 대하여 표본을 얻기도 하지만 처음부터 두 그룹이 독립적으로 구성이 되어 있는 경우가 있다. 두 나라의 평균 소득을 비교하고자 할 때 각 나라에서 모든 국민의 소득이 두 모집단을 이루며 이로부터 조건에 따라 표본을 추출하면 이는 독립표본이 된다.

수집된 자료인 표본을 통해 두 집단을 비교하기 위해선 통계량은 각 집단의 평균 또는 비율을 가지고 하게 되며 이들 통계량은 두 모집단의 특성을 나타내는 모수를 비교하는 것 대신에 이용된다. 특히 7장에서 설명한 바와 같이 모집단의 모수로 모평균 μ과 모비율 p를 사용하며 두 집단을 비교하기 위해서는 다음과 같은 가설이 필요하다.

■ 대응표본의 귀무가설

H_o : 한 모집단의 두 번 측정된 모평균의 차이는 없다. $\Leftrightarrow H_o : \mu_1 - \mu_2 = 0$

■ 독립표본의 귀무가설

H_o : 두 모집단의 평균(모평균)은 차이가 없다. $\Leftrightarrow H_o : \mu_1 - \mu_2 = 0$
H_o : 두 모집단의 비율(모비율)은 차이가 없다. $\Leftrightarrow H_o : p_1 - p_2 = 0$

8.2 독립표본의 비교

두 독립표본의 비교를 위하여 이들이 독립적으로 표집된 두 모집단을 X와 Y라 했을 때 이들과 관련된 모수, 표본의 크기, 표본, 통계량과 같은 정보는 다음과 같이 요약할 수 있다.

모집단	X	Y
모수(모평균과 모분산)	$\mu_1,\ \sigma_1^2$	$\mu_2,\ \sigma_2^2$
표본의 크기	n_1	n_2
표본	$X_1, X_2,\ \ldots, X_{n_1}$	$Y_1, Y_2,\ \ldots, Y_{n_2}$
통계량(평균과 분산)	$\overline{X}=\frac{1}{n_1}\sum_{i=1}^{n_1}X_i$ $s_1^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\overline{X})^2$	$\overline{Y}=\frac{1}{n_2}\sum_{i=1}^{n_2}Y_i$ $s_2^2=\frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2$

일반적으로 표본은 크기에 따라 30보다 작으면 소표본이고 30보다 크면 대표본으로 나누어진다. 소표본의 경우 두 모집단의 가정이 모평균과 모분산을 가진 정규분포인 $X \sim N(\mu_1, \sigma_1^2)$와 $Y \sim N(\mu_2, \sigma_2^2)$가 만족되어야 하며 이로부터 얻어진 두 독립표본의 평균들도 다음과 같이 정규분포를 만족한다.

$$\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1}X_i \sim N\left(\mu_1,\ \frac{\sigma_1^2}{n_1}\right),\ \overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2}Y_i \sim N\left(\mu_2,\ \frac{\sigma_2^2}{n_2}\right) \tag{8.1}$$

대표본의 경우에는 두 모집단에 대한 모평균과 모분산에 대한 가정만 있고 정규분포 가정이 없는 $X \sim (\mu_1, \sigma_1^2)$와 $Y \sim (\mu_2, \sigma_2^2)$일지라도 두 독립표본의 평균들은 중심극한정리(central limit theorem)에 의해서 근사적으로 정규분포를 만족하게 된다.

따라서 표본의 크기에 따라 소표본과 대표본으로 나누어지고 이들이 표집된 모집단에 대한 기본 가정도 다르며 이에 따라 두 독립표본을 비교하는 검정통계량도 달라진다. 다만 비교하고자 하는 귀무가설과 검정절차는 동일하다는 점이다.

우선 8.1절에서 독립표본에 대하여 언급한 다음의 귀무가설

$$H_o : \text{두 모집단의 평균(모평균)은 차이가 없다.} \Leftrightarrow H_o : \mu_1 - \mu_2 = 0$$

에서 보고자 하는 모평균의 차이 $\mu_1 - \mu_2$에 대하여 사용하는 검정통계량의 바탕이 되는 통계량은 두 독립표본의 평균의 차 $\overline{X} - \overline{Y}$를 이용하며 이것의 평균과 분산은 다음과 같다.

$$\overline{X} - \overline{Y} \sim \left(\mu_1 - \mu_2,\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \tag{8.2}$$

이 된다. 이를 바탕으로 검정통계량을 소표본과 대표본에 따라 구성하여 보자.

소표본의 경우에는 두 독립표본이 정규분포를 반드시 만족하여 식 (8.1)이 성립하고 만약에 $\sigma_1^2 = \sigma_1^2 = \sigma^2$으로 두 모분산이 동일하다는 가정이 성립하면 식 (8.2)는

$$\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2,\ \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right) \tag{8.3}$$

가 된다. 식 (8.3)으로부터 $\overline{X} - \overline{Y}$를 표준화하면 통계량

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim N(0,\ 1) \tag{8.4}$$

이 된다. 일반적으로 식 (8.4)에서 모분산 σ^2은 잘 알려져 있지 않고 모르는 경우 표본분산 s_1^2과 s_2^2에 의해서 추정하여 합동표본분산(pooled sample variance)

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

을 사용하게 되며 식 (8.4)의 표준정규분포는 성립하지 않게 된다. 따라서 이 경우 다음과 같은 통계량

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{df}$$

를 얻는데 이는 자유도 $df = n_1 + n_2 - 2$인 t-분포를 따른다. 따라서 귀무가설 H_o : $\mu_1 - \mu_2 = 0$하에 검정통계량은 다음과 같다.

$$T = \frac{\overline{X} - \overline{Y}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}. \tag{8.5}$$

검정방법은 7장에서 언급한 바와 같이 유의수준 α가 주어지면 자유도가 $n_1 + n_2 - 2$인 t-분포표로부터 대립가설의 형태에 따라 단측이면 t_α, 양측이면 $t_{\alpha/2}$를 기각치로 하며 이는 식 (8.5)의 통계량의 관측값 t가 다음의 기각역을 만족하면 귀무가설 H_o : $\mu_1 - \mu_2 = 0$을 기각한다.

■ 소표본에서 독립표본의 두 모평균에 대한 $t-$검정

귀무가설	대립가설	기각역
	$H_1 : \mu_1 - \mu_2 > 0$	$t \geqq t_\alpha$
$H_0 : \mu_1 - \mu_2 = 0$	$H_1 : \mu_1 - \mu_2 < 0$	$t \leqq -t_\alpha$
	$H_1 : \mu_1 - \mu_2 \neq 0$	$\lvert t \rvert \geqq t_{\alpha/2}$

물론 식 (8.4)의 검정통계량의 관측값 t에 의해 대립가설의 형태에 따라 각각 구한 p-값인 $P(T \geqq t)$, $P(T \leqq t)$, $P(\lvert T \rvert \geqq t)=2P(T \geqq t)=2P(T \leqq -t)$에 의해서 유의수준 α보다 작으면 귀무가설을 기각하는 검정이 가능하다. 다만 7장에서 언급한 바와 같이 분포표를 이용하여 정확하게 이 확률값을 구하기가 쉽지 않다. 본문 중 보기에서 이들 p-값 대신에 기각역을 이용한 검정을 다루고자 한다. 다만 8.6절에는 R-프로그램을 활용한 검정에서는 기각역에 의한 것보다 p-값에 의한 검정을 제공하고 있으므로 참고하기 바란다.

보기 8.1

연령이 7세에서 11세 남녀 어린이가 하루 중 운동을 하는 데 보내는 평균시간이 같다고 알려져 있다. 그러나 새로운 연구 결과 평균시간이 서로 다르다고 하여 모분산이 같고 정규분포를 따르는 모집단으로부터 남자 9명과 여자 16명을 추출하여 조사하였더니 다음과 같았다. 이를 토대로 유의수준 5%에서 남녀 집단 간의 평균시간의 차에 대한 검정을 해보라.

모집단	표본의 크기	평균시간	표준편차
남자 X	16	3.2	1.00
여자 Y	9	2	$\sqrt{0.75}$

1) 남녀 집단의 모평균을 각각 μ_1과 μ_2라 하면 다음과 같은 가설을 세울 수 있다.

$$H_o : \mu_1 - \mu_2 = 0, \quad H_1 : \mu_1 - \mu_2 \neq 0$$

2) 표본의 크기 16과 9가 30보다 작아 소표본에 해당하고 모분산이 동일하며 정규분포를 따

르므로 합동표본분산.

$$s_p^2 = \frac{(16-1)1.0+(9-1)0.75}{16+9-2} = 0.75$$

을 계산한다. 따라서 검정통계량의 관측값은 다음과 같다.

$$t = \frac{3.2-2}{0.87\sqrt{\frac{1}{16}+\frac{1}{9}}} = 4.06$$

3) 대립가설이 양측인 $H_1 : \mu_1 - \mu_2 \neq 0$이므로 유의수준 5%에서 자유도 $df = n_1 + n_2 - 2 = 23$인 기각치는 $t_{\alpha/2} = t_{0.025} = 2.069$이므로 따라서 기각역은 $|t| \geqq 2.069$이다.

4) 검정통계량의 관측값 $t = 4.06$의 절대값이 기각역을 만족하므로 귀무가설 H_o : $\mu_1 - \mu_2 = 0$을 기각하게 된다.

5) 이 연구에 따르자면 남녀 간에 하루 중 운동을 하는 데 보내는 평균시간은 차이가 있다고 할 수 있다.

보기 8.2

젖소용 사료 X와 Y의 우유생산량이 같다는 연구결과가 있다. 이를 입증하기 위하여 정규분포를 따르며 분산이 동일한 두 모집단으로부터 각각 추출한 13마리에는 사료 X를 공급하고 나머지 12마리에는 사료 Y를 공급하였다. 3주일 후 젖소의 우유 생산량 검사를 해본 결과 다음의 자료를 얻었다. 알려진 연구 결과와는 달리 사료 X의 우유생산량이 사료 Y의 생산량보다 더 많다고 할 수 있는지를 유의수준 5%에서 검정하라.

모집단	자료 크기	평균	표준편차
사료 X	$n_1 = 13$	45.15	7.998
사료 Y	$n_2 = 12$	42.25	8.740

1) 사료 X와 Y의 우유생산량의 모평균을 각각 μ_1과 μ_2라 하면 다음과 같은 가설을 세울 수 있다.

$$H_o: \mu_1 - \mu_2 = 0, \quad H_1: \mu_1 - \mu_2 > 0$$

2) 표본의 크기가 13과 12로 소표본에 해당하고 모분산이 동일하며 정규분포를 따르므로 합동표본분산에 의한

$$s_p = \sqrt{\frac{12(7.998)^2 + 11(8.740)^2}{23}} = 8.36$$

을 계산한다. 따라서 검정통계량의 관측값은 다음과 같다.

$$t = \frac{45.15 - 42.25}{8.36\sqrt{\frac{1}{13} + \frac{1}{12}}} = \frac{2.90}{3.347} = 0.87$$

3) 대립가설이 단측인 $H_1: \mu_1 - \mu_2 > 0$이므로 유의수준 5%에서 자유도 $df = n_1 + n_2 - 2 = 23$인 기각치는 $t_{.05} = 1.714$이고 기각역은 $t \geq 1.714$이다.

4) 검정통계량의 관측값 $t = 0.87$이 기각역을 벗어나므로 귀무가설 $H_o: \mu_1 - \mu_2 = 0$을 기각하지 못한다.

5) 사료 X의 우유생산량이 사료 Y의 생산량보다 더 많다고 할 수 없다.

대표본의 경우에는 두 독립표본이 정규분포를 만족하지 않더라도 근사적으로 식 (8.1)을 만족하게 되고 모분산이 동일하다는 가정을 필요치는 않으며 두 독립표본의 평균의 차 $\overline{X} - \overline{Y}$는 근사적으로 정규분포를 다음과 같이 따른다.

$$\overline{X} - \overline{Y} \simeq N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \tag{8.6}$$

식 (8.6)에서 두 분산을 아는 경우에는 $\overline{X} - \overline{Y}$를 표준화하면 근사적으로 표준정규분포를 따르는 통계량

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \simeq N(0,\ 1) \tag{8.7}$$

를 얻게 된다. 만약에 식 (8.6)에서 두 모분산을 모르는 경우 각 분산의 추정량인 표본분산 s_1^2과 s_2^2를 사용해도 대표본이기 때문에 다음이 성립한다.

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \simeq N(0,\ 1) \tag{8.8}$$

따라서 귀무가설 $H_o : \mu_1 - \mu_2 = 0$하에 식 (8.7)과 식 (8.8)로부터 검정통계량은 다음과 같다.

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad \text{또는} \quad Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{8.9}$$

검정방법은 유의수준 α가 주어지면 표준정규분포표로부터 대립가설의 형태에 따라 단측이면 z_α, 양측이면 $z_{\alpha/2}$를 기각치로 각각 사용한다. 다시 말해서 식 (8.8)의 통계량의 관측값 z에 의해서 대립가설에 따라 기각치와의 관계가 다음과 같은 기각역을 만족하면 귀무가설 H_o : $\mu_1 - \mu_2 = 0$을 기각한다.

■ 대표본에서 독립표본의 두 모평균에 대한 $t-$검정

귀무가설	대립가설	기각역
	$H_1 : \mu_1 - \mu_2 > 0$	$z \geqq z_\alpha$
$H_0 : \mu_1 - \mu_2 = 0$	$H_1 : \mu_1 - \mu_2 < 0$	$z \leqq -z_\alpha$
	$H_1 : \mu_1 - \mu_2 \neq 0$	$\lvert z \rvert \geqq z_{\alpha/2}$

물론 식 (8.8)의 검정통계량의 관측값 z에 의해 대립가설의 형태에 따라 각각 구한 p-값인 $P(Z \geqq z)$, $P(Z \leqq z)$, $P(\lvert Z \rvert \geqq z)$=$2P(Z \geqq z)$=$2P(Z \leqq -z)$에 의해서도 검정이 가능하다. 이를 활용한 검정은 8.6절의 [프로그램 8.1]과 [결과 8.1]을 참고 바란다.

보기 8.3

두 기업 중 노동조합의 활동이 활발한 기업이 그렇지 않은 기업보다 근로자의 평균임금이 더 높다는 보고가 있다. 이를 검정하기 위하여 노동조합이 활발한 기업으로부터 50명과 노동조합이 활발하지 않은 기업으로부터 30명을 임의로 선발하여 임금을 조사한 결과 다음의 자료를 얻었다. 알려진 보고 결과와 같이 근로자들의 평균임금이 노동조합이 활발한 기업 X가 노동조합이 활발하지 않은 기업 Y보다 더 높다고 할 수 있는 지를 유의수준 5%에서 검정하라.

모집단	표본의 크기	평균임금(원)	표준편차
기업 X	50	225,200	9,800
기업 Y	30	218,700	10,800

1) 기업 X와 Y의 근로자 임금의 모평균을 각각 μ_1과 μ_2라 하면 다음과 같은 가설을 세울 수 있다.

$$H_o : \mu_1 - \mu_2 = 0, \quad H_1 : \mu_1 - \mu_2 > 0$$

2) 표본의 크기 50과 30은 대표본에 해당하며 검정통계량의 관측값 다음과 같다.

$$z = \frac{225200 - 218700}{\sqrt{\frac{(9800)^2}{50} + \frac{(10800)^2}{30}}} = 2.70$$

3) 대립가설 $H_1 : \mu_1 - \mu_2 > 0$에서 유의수준 5%의 기각치는 $z_{.05} = 1.64$로 기각역은 $z \geq 1.64$이다. 검정통계량의 관측값 $z = 2.70$이 기각역을 만족하므로 귀무가설 H_o : $\mu_1 - \mu_2 = 0$을 기각하게 된다.

4) 근로자들의 평균임금은 노동조합이 활발한 기업 X가 노동조합이 활발하지 않은 기업 Y보다 더 높다고 할 수 있다.

지금까지 독립표본 모평균 차의 검정 절차를 소표본과 대표본에 따라 요약 정리한 것이 [표 8.1]이다.

[표 8.1] 소표본과 대표본에 따른 독립표본 모평균 차의 검정 절차

표본의 크기	소표본(n_1, n_2 ≤ 30)	대표본(n_1, n_2 > 30)
모집단의 정규성 가정	$X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$	$X \sim (\mu_1, \sigma_1^2)$, $Y \sim (\mu_2, \sigma_2^2)$
모르는 모분산 가정	$\sigma_1^2 = \sigma_1^2 = \sigma^2$	$\sigma_1^2 \neq \sigma_1^2$
검정통계량	$T = \dfrac{\overline{X} - \overline{Y}}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$, s_p^2 : 합동표본분산	$Z = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$, s_1^2, s_2^2 : 표본분산
대립가설과 기각역 (유의수준 α)	$H_1 : \mu_1 - \mu_2 > 0$, $t \geq t_\alpha$, $H_1 : \mu_1 - \mu_2 < 0$, $t \leq -t_\alpha$, $H_1 : \mu_1 - \mu_2 \neq 0$, $\lvert t \rvert \geq t_{\alpha/2}$.	$H_1 : \mu_1 - \mu_2 > 0$, $z \geq z_\alpha$, $H_1 : \mu_1 - \mu_2 < 0$, $z \leq -z_\alpha$, $H_1 : \mu_1 - \mu_2 \neq 0$, $\lvert z \rvert \geq z_{\alpha/2}$.
$p-$값	$P(T \geq t)$, $P(T \leq t)$, $P(\lvert T \rvert \geq t)$	$P(Z \geq z)$, $P(Z \leq z)$, $P(\lvert Z \rvert \geq z)$
검정방법	검정통계량의 관측값이 기각역을 만족하든지 $p-$값$< \alpha$이면 귀무가설 $H_o : \mu_1 - \mu_2 = 0$을 기각한다.	

8.3 소표본에서 모분산이 다른 경우의 비교

두 독립표본을 비교하는 가설과 검정에서 표본의 크기가 30보다 큰 대표본의 경우에는 이들이 추출된 두 모집단 $X\sim(\mu_1,\ \sigma_1^2)$와 $Y\sim(\mu_2,\ \sigma_2^2)$에 대해 정규분포 가정이 없고 두 분산이 같다는 가정이 없어도 별 문제가 없다. 다만 표본의 크기가 30보다 작은 소표본의 경우 8.2절에서 살펴본 것처럼 반드시 다음의 가정이 요구된다.

- 두 모집단이 정규분포 $X\sim N(\mu_1,\ \sigma_1^2)$와 $Y\sim N(\mu_2,\ \sigma_2^2)$를 따른다.
- 두 모집단의 모분산은 $\sigma_1^2=\sigma_1^2=\sigma^2$으로 동일하며 알려져 있지 않다.
- 두 모집단으로부터 추출된 표본은 서로 독립이다.

여기서 제기하는 문제는 두 모분산이 같다는 $\sigma_1^2=\sigma_1^2=\sigma^2$에 대한 가정이 성립하지 않은 경우 합동표본분산을 사용하는 식 (8.5)의 자유도 $df=n_1+n_2-2$인 t-분포를 따르는 검정통계량을 사용할 수 없다는 점이다. 대신에 대표본에서 사용하는 검정통계량을 사용하되 이는 다음과 같이 근사적으로 t-분포를 따른다.

$$T=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}}}\simeq t_{df}$$

여기서 웰치(Welch)의 수정된 자유도는

$$df=\frac{\left[\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}\right]^2}{\dfrac{1}{n_1-1}\left[\dfrac{s_1^2}{n_1}\right]^2+\dfrac{1}{n_2-1}\left[\dfrac{s_2^2}{n_2}\right]^2} \tag{8.10}$$

이다. 만약에 표본으로부터 계산된 식 (8.10)의 자유도 값이 정수가 되지 않으면 반올림 한 후 소수값을 제외하고 정수값만 사용해도 된다. 계산상 더 편리하게 $n_1 - 1$과 $n_2 - 1$중 작은 값을 자유도로 정하는 방법도 있다. 실제로 표본으로부터 두 표본표준편차 s_1과 s_2의 상대적 크기가 $0.5 \le s_1/s_2 \le 2$일 때 두 모분산이 같다는 $\sigma_1^2 = \sigma_2^2$ 가정을 할 수 있다.

보기 8.4

새롭게 개발된 농약이 과실의 수확량을 높인다는 연구가 있다. 실제로 과수원에서 9그루 나무에 기존의 농약을 사용하였고, 나머지 9그루 나무에 새 농약을 사용하여 수확한 과실의 수확량에 대한 다음의 자료를 얻었다. 연구결과와 같이 새로운 농약 X가 기존의 농약 Y보다 과실의 평균 수확량을 더 많이 낸다고 할 수 있는 지를 유의수준 1%에서 검정하라.

모집단	표본의 크기	평균 수확량	표준편차
농약 X	9	249	19
농약 Y	9	233	45

1) 새 농약 X와 기존의 농약 Y의 수확량의 모평균을 각각 μ_1과 μ_2라 하면 다음과 같은 가설을 세울 수 있다.

$$H_o : \mu_1 - \mu_2 = 0, \quad H_1 : \mu_1 - \mu_2 > 0$$

2) 두 표본의 크기 9는 소표본에 해당하며 두 표준편차의 비가 $19/45 = 0.42$로 두 모평균이 같다고 보기 어렵다. 따라서 수정된 자유도

$$df = \frac{\left[\frac{19^2}{9} + \frac{45^2}{9}\right]^2}{\frac{1}{8}\left[\frac{19^2}{9}\right]^2 + \frac{1}{8}\left[\frac{45^2}{9}\right]^2} = 10.765$$

계산으로부터 자유도는 $df = 11$이고 검정통계량의 관측값 다음과 같다.

$$t = \frac{249 - 233}{\sqrt{\frac{19^2}{9} + \frac{(45)^2}{9}}} = 9.98$$

3) 대립가설이 단측인 $H_1 : \mu_1 - \mu_2 > 0$에 대해 유의수준 1%에서 자유도 $df = 11$인 기각치 $t_{.01} = 2.718$이므로 기각역은 $t \geq 2.718$이다. 검정통계량의 관측값 $t = 9.98$가 기각역을 만족하므로 귀무가설 $H_o : \mu_1 - \mu_2 = 0$을 기각한다.

4) 새로 개발된 농약 X가 기존 농약 Y의 농작물 평균 수확량보다 더 많이 낸다고 할 수 있다.

만약에 [보기 8.4]의 풀이 2)에서 수정된 자유도 계산 대신에 두 표본의 크기로부터 $n_1 - 1$과 $n_2 - 1$중 작은 값을 자유도로 정하는 방법에 따르면 두 표본의 크기가 9로 같으므로 자유도는 8이 된다. 이 때 기각치는 $t_{.01} = 2.896$이며 검정의 결과는 동일하다.

8.4 대응표본

8.3절에서는 남녀 어린이 하루 중 운동을 하는 데 보내는 평균시간, 젖소용 두 사료의 우유 생산량, 두 기업 중 노동조합의 활동이 활발한 기업과 활발하지 않은 기업 근로자의 평균임금, 새롭게 개발된 농약과 기존 농약의 과실 수확량과 같이 두 그룹은 서로에게 영향을 끼칠 수 없는 독립표본이 된다. 물론 실험하기 전 두 그룹의 여러 조건(남녀 어린이의 연령, 평소 젖소의 우유 생산량, 회사의 규모, 과실나무의 상태)이 유사하도록 확보되어야 함은 당연하다.

그러나 실험에 따라 두 그룹의 유사한 조건을 보장하기 어렵거나 충분한 대상을 확보하지 못하는 경우 대안으로 동일한 대상에 대하여 시간을 두고 측정하는 대응표본을 얻게 된다. 이 경우 먼저 실시하는 실험의 효과가 남아 있지 않도록 주의해야 한다. 다행이 무좀과 같은 피부 치료제의 경우에는 무좀 환자의 한쪽 발에는 새로운 약을 다른 한쪽 발에는 기존의 약을 바르고 결과를 관찰하거나, 두 종류의 자동차 타이어 수명을 비교하고자 할 때 한 자동차에 두 종류 타이어를 한꺼번에 부착하여 주행 실험한다면 이러한 문제는 쉽게 해결된다.

일반적으로 n개의 실험대상에 대하여 두 번 처리한 것을 각각 처리 1과 처리 2로 하고 i번째 대상에 대한 각 처리를 X_i와 Y_i라 하면 두 처리간의 차이 $D_i = X_i - Y_i$, $i = 1, \ldots, n$는 표본평균 $\overline{D}$와 표본분산 s_D^2를 갖는 새로운 표본인 대응표본을 이루며 다음과 같이 요약할 수 있다.

	처리 대상					
	1	2	$\cdots$	i	$\cdots$	n
처리 1	X_1	X_2	$\cdots$	X_i	$\cdots$	X_n
처리 2	Y_1	Y_2	$\cdots$	Y_i	$\cdots$	Y_n
차	$D_1 = X_1 - Y_1$	$D_2 = X_2 - Y_2$	$\cdots$	$D_i = X_i - Y_i$	$\cdots$	$D_n = X_n - Y_n$
통계량	새로운 자료 $D_1, D_2, \cdots, D_n$의 표본평균 $\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$, 표본분산 $s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2$					

여기서 두 처리 간의 차에 의해 구해진 새로운 자료 $D_1, D_2, \cdots, D_n$은 모평균 μ_D와 알려져 있지 않은 모분산 σ_D^2를 따른다고 하자. 특히, 소표본의 경우에는 $D_1, D_2, \cdots, D_n$가 정규분포 $N(\mu_D, \sigma_D^2)$를 따르며 이 경우 두 처리 간의 차이의 모평균 μ_D에 대한 귀무가설 $H_0 : \mu_D = 0$에 대한 검정통계량은 7.4절의 한 모집단 모평균에 대한 검정통계량과 비슷하게 자유도 $df = n-1$인 t-분포를 따르는 다음의 통계량에 바탕을 둔다.

$$T = \frac{\overline{D} - \mu_D}{s_D / \sqrt{n}} \sim t_{df} \tag{8.11}$$

귀무가설 $H_o : \mu_D = 0$하에 검정통계량은 다음과 같다.

$$T = \frac{\overline{D}}{s_D / \sqrt{n}} \tag{8.12}$$

검정방법은 대립가설의 형태에 따른 기각역과 $p-$값을 활용하는 일반적인 방법을 따르므로 생략한다.

만약에 표본의 크기가 대표본이면 식 (8.11)의 통계량은 t-분포 대신에 표준정규분포를 따르게 된다. 다만 귀무가설 $H_o : \mu_D = 0$을 검정하기 위한 검정통계량의 형태는 식 (8.12)와 동일하며 자유도가 무한대 ∞인 t-분포를 사용해도 표준정규분포와 비교할 때 검정의 결과는 별 차이가 없다.

보기 8.5

두 아웃도어 회사의 등산화 밑창의 마모 정도를 비교하기 위하여 10명의 학생들을 대상으로 X회사 제품을 왼발에 착용하였고 오른발에는 Y회사 제품을 각각 착용하게 하여 마모를 측정한 결과 다음의 자료를 얻었다. X회사 제품의 등산화의 마모율이 Y회사 제품과 차이가 있다고 할 수 있는 지를 유의수준 5%에서 검정하라.

모집단	처리 대상									
제품 X (왼발)	14.0	8.8	11.2	14.2	11.8	6.4	9.8	11.3	9.3	13.6
제품 Y (오른발)	13.2	8.2	10.9	14.3	10.7	6.6	9.5	10.8	8.8	13.3
차 D	0.8	0.6	0.3	−0.1	1.1	−0.2	0.3	0.5	0.5	0.3

1) 제품 X와 제품 Y의 마모율 차이의 모평균 μ_D에 대한 귀무가설과 대립가설은 다음과 같다.

$$H_0 : \mu_D = 0,\ H_1 : \mu_D \neq 0$$

2) 대응표본의 크기 10은 소표본에 해당하며 따라서 표본평균과 표본분산의

$$\overline{D} = \frac{1}{10}\sum_{i=1}^{n} D_i = 0.41, \qquad s_D^2 = \frac{1}{10-1}\sum_{i=1}^{n}(D_i - \overline{D})^2 = 0.0149$$

계산으로부터 귀무가설 $H_o : \mu_D = 0$하에 검정통계량의 관측값은 다음과 같다.

$$t = \frac{0.41}{0.122/\sqrt{10}} = 3.349$$

3) 대립가설이 양측인 $H_1 : \mu_D \neq 0$에 대해 유의수준 5%에서 자유도 $df = n-1 = 9$에서 기각치 $t_{.0.025} = 2.262$에 의해서 기각역은 $|t| \geq 2.262$이다. 검정통계량의 관측값의 $|t| = 3.349$가 기각역을 만족하므로 귀무가설 $H_o : \mu_D = 0$을 기각한다.

4) 등산화 제품 X와 제품 Y의 마모율은 차이가 있다고 볼 수 있다.

보기 8.6

어느 TV 방송국에서 봄철 프로그램 개편 전과 후 TV 시청률의 차이를 보기 위해 시청률 조사기관에 의뢰하여 400세대를 대상으로 하루 동안의 시청시간을 조사한 다음의 자료를 받았다. 개편 전의 시청시간을 X와 개편 후의 시청시간 Y의 차를 $D = X - Y$라 하면 자료의 결과는 다음과 같다. 개편 후의 시청시간 Y가 개편 전의 시청시간 X보다 늘었는지를 유의수준 1%에서 검정하라.

모집단	표본의 크기	차 평균	표준편차
개편 전 X	400	- 0.21	2.23
개편 후 Y			

1) 개편 전 시청시간 X와 개편 후 시청시간 Y의 차 $D = X - Y$의 모평균 μ_D에 대한 귀무가설과 대립가설을 다음과 같이 세울 수 있다.

$$H_0 : \mu_D = 0, \; H_1 : \mu_D < 0$$

2) 표본평균과 표본분산

$$\overline{D} = -0.21, \; s_D^2 = 2.23^2$$

으로부터 귀무가설 $H_o : \mu_D = 0$하에 검정통계량의 관측값은 다음과 같다.

$$t = \frac{-0.21}{2.23/\sqrt{400}} = -1.88$$

3) 대립가설 $H_1 : \mu_D < 0$에 대해 유의수준 1%에서 대응표본의 크기가 400으로 대표본에 해당하므로 자유도 $df = \infty$에서 기각치 $-t_{.0.01} = -2.326$에 의해 기각역은 $t \leq -2.326$이다. 따라서 검정통계량의 관측값 $t = -1.88$이 기각역을 벗어나므로 귀무가설 $H_0 : \mu_D = 0$을 기각하지 못한다.

4) 봄철 프로그램 개편 후의 시청시간이 늘었다는 주장을 받아들일 수 없다.

8.5 독립표본의 모비율 비교

지금까지 독립표본과 대응표본의 모평균 μ에 대한 비교를 설명하였다. 이 절에서는 독립표본에서 모평균 대신 모비율 p를 비교하는 가설 검정을 다루고자 한다. 8.1절에서 설명한 바와 같이 아스피린이 심장발작을 예방한다는 효과에 대해 아스피린을 복용한 그룹과 위약(placebo)을 복용한 그룹으로 나누어 여러 해 동안 두 그룹의 심장발작 비율을 비교하거나, 부산시에서 남구와 서구의 노동력 인구와 이들 중 실업자의 비율, 그리고 제안된 정책에 대하여 남녀 간의 의견차에 대한 비율을 비교를 하고자 한다.

먼저 모비율 p_1과 p_2인 두 모집단으로부터 추출된 크기가 n_1과 n_2인 두 독립표본으로부터 X와 Y를 성공의 횟수라 하자. 궁극적인 두 그룹의 모비율에 대한 비교를 위해선 다음과 같은 귀무가설이 요구된다.

$$H_o : \text{두 모집단의 비율(모비율)은 차이가 없다.} \Leftrightarrow H_o : p_1 - p_2 = 0$$

두 독립표본의 크기가 n_1과 n_2가 큰 대표본인 경우 모비율 p_1과 p_2의 추정량은 근사적으로 정규분포를 따른다.

$$\hat{p_1} = \frac{X}{n_1} \simeq N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right), \quad \hat{p_2} = \frac{Y}{n_2} \simeq N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right) \tag{8.13}$$

두 모집단의 비율(모비율)의 차에 대한 귀무가설 $H_o : p_1 - p_2 = 0$을 위한 검정통계량의 바탕이 되는 통계량은 $p_1 - p_2$의 추정치 $\hat{p_1} - \hat{p_2}$가 되며 이는 식 (8.13)으로부터 근사적으로 정규분포

$$\hat{p_1} - \hat{p_2} \simeq N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_1(1-p_1)}{n_1}\right)$$

를 따르며 이를 표준화하면 다음과 같이 근사적으로 표준정규분포를 따른다.

$$Z = \frac{(\hat{p_1} - \hat{p_2}) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \simeq N(0,1) \qquad (8.14)$$

식 (8.14)에서 귀무가설 $H_o:\ p_1 - p_2 = 0$하에 모비율 p_1과 p_2는 $p_1 = p_2 = p$를 만족하고 이를 잘 모르나 두 독립표본이 대표본이므로 이들의 합동 추정치

$$\hat{p} = \frac{X+Y}{n_1+n_2} = \frac{n_1\hat{p_1} + n_2\hat{p_2}}{n_1+n_2}$$

를 사용할 수 있으며 이것에 의해서 다음과 같은 검정통계량을 이용할 수 있다.

$$Z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (8.15)$$

검정방법은 유의수준 α가 주어지면 표준정규분포표로부터 대립가설의 형태에 따라 단측이면 z_α, 양측이면 $z_{\alpha/2}$를 기각치로 각각 사용한다. 다시 말해서 식 (8.15)의 통계량의 관측값 z에 의해서 대립가설에 따라 기각치와의 관계가 다음과 같은 기각역을 만족하면 귀무가설 $H_o:\ p_1 - p_2 = 0$을 기각한다.

■ 독립표본의 두 모비율에 대한 유의수준 α에서 $Z-$검정

귀무가설	대립가설	기각역
	$H_1 : p_1 - p_2 > 0$	$z \geqq z_\alpha$
$H_0 : p_1 - p_2 = 0$	$H_1 : p_1 - p_2 < 0$	$z \leqq -z_\alpha$
	$H_1 : p_1 - p_2 \neq 0$	$\lvert z \rvert \geqq z_{\alpha/2}$

더불어 식 (8.15)의 검정통계량의 관측값 z에 의해 대립가설의 형태에 따라 각각 구한 p-값인 $P(Z \geqq z)$, $P(Z \leqq z)$, $P(\lvert Z \rvert \geqq z)=2P(Z \geqq -z)=2P(Z \leqq z)$에 의해서도 검정이 가능하다. 이를 활용한 검정은 8.6절의 [프로그램 8.3]과 [결과 8.3]을 참고 바란다.

보기 8.7

독감 예방 백신을 60세에서 65세 노인들에게 접종을 하였다. 남성 노인 113명 가운데 34명이 항체가 생겼고 여성 노인 139명 가운데 54명이 항체가 생겼다. 항체의 생성 비율이 남성 노인보다 여성 노인 쪽이 더 높다고 할 수 있는지를 유의수준 5%에서 검정하고 $p-$값에 의한 결과와 비교하라.

모집단	표본의 크기	항체 생성
남자	113	$X = 34$
여자	139	$Y = 54$

1) 남성 노인과 여성 노인의 항체 생성 모비율을 각각 p_1과 p_2 하면 다음과 같은 가설을 세울 수 있다.

$$H_o : p_1 - p_2 = 0, \quad H_1 : p_1 - p_2 < 0$$

2) 두 독립표본의 크기 $n_1 = 113$과 $n_2 = 139$는 매우 큰 값으로 대표본에 해당하며 합동 추정치 $\hat{p} = \dfrac{34+54}{113+139} = 0.349$으로부터 검정통계량의 관측값은 다음과 같이 계산된다.

$$z = \frac{0.301 - 0.388}{\sqrt{0.349 \times (1 - 0.349)}\sqrt{\frac{1}{113} + \frac{1}{139}}} = -1.45$$

3) 대립가설 $H_1 : p_1 - p_2 < 0$에 대하여 유의수준 $5\% = 0.05$에서 기각치는 $-z_{.05} = -1.64$로 기각역은 $z \leqq -1.64$이다. 검정통계량의 관측값 $z = -1.45$는 기각역을 벗어나므로 귀무가설 $H_o : p_1 - p_2 = 0$을 기각하지 못하게 된다.

4) 여성 노인의 항체 생성률이 남성 노인의 항체 생성률보다 높다고 할 수 없다.

보기 8.8

부산광역시 남구와 해운대구에 거주하는 주민의 자기주택보유율이 다른지를 보기 위해 남구에서 540명과 해운대구에서 432명을 대상으로 자기 집 소유에 대한 설문조사를 하였다. 이 결과 두 거주지의 자기주택보유율이 다른지를 유의수준 5%에서 검정하고 p−값에 의한 결과와 비교하라.

모집단	표본의 크기	자기 집 소유
남구	540	X = 362
해운대구	432	Y = 281

1) 남성 노인과 여성 노인의 항체 생성 모비율을 각각 p_1과 p_2 하면 다음과 같은 가설을 세울 수 있다.

$$H_o : p_1 - p_2 = 0, \quad H_1 : p_1 - p_2 < 0$$

2) 두 독립표본의 크기 $n_1 = 113$과 $n_2 = 139$는 매우 큰 값으로 대표본에 해당하며 합동 추정치 $\hat{p} = \frac{34 + 54}{113 + 139} = 0.349$으로부터 검정통계량의 관측값은 다음과 같이 계산된다.

$$z = \frac{0.301 - 0.388}{\sqrt{0.349 \times (1 - 0.349)} \sqrt{\frac{1}{113} + \frac{1}{139}}} = -1.45$$

3) 대립가설 $H_1 : p_1 - p_2 < 0$에 대하여 유의수준 $5\% = 0.05$에서 기각치는 $-z_{.05} = -1.64$로 기각역은 $z \leq -1.64$이다. 검정통계량의 관측값 $z = -1.45$는 기각역을 벗어나므로 귀무가설 $H_o : p_1 - p_2 = 0$을 기각하지 못하게 된다.

4) 여성 노인의 항체 생성률이 남성 노인의 항체 생성률보다 높다고 할 수 없다.

8.6 R - 프로그램 실습

[프로그램 8.1]은 독립표본의 두 모평균 검정의 예로 [보기 8.2]에서 젖소용 사료 X와 Y의 우유생산량이 같다는 연구 가설 $H_o : \mu_1 - \mu_2 = 0$, $H_1 : \mu_1 - \mu_2 > 0$을 검정하기 위한 t.test( ) 함수를 활용한 R - 프로그램이다. 이 함수를 사용하기 위해서 독립된 두 표본의 자료 <x=c(44, 44, …, 41), y=c(35, 47, …, 39)>가 필요하다. 함수에서 옵션으로 두 표본의 분산이 같다는 가정을 위해 <var.equal=T> 옵션을 사용하며 7장에서 언급한 바와 같이 대립가설의 형태인 양측과 단측에 따라 <alt(ernative) = "two.sided", “ less", "greater">를 설정한다. 참고로 8.3절의 소표본에서 모분산이 다른 경우 웰치(Welch)의 수정된 자유도를 사용하는 검정은 <var.equal=F> 옵션을 사용하면 된다. 검정방법으로 기각역을 활용하는 것 대신에 유의확률인 $p-$값을 제시하고 있다. 여기서 <$p-$값 = $P(T \geq 0.87)$=0.1973>은 유의수준 $\alpha = 0.05$에 비해 크므로 젖소용 사료 X와 Y의 우유생산량이 같다는 연구 가설 H_o : $\mu_1 - \mu_2 = 0$을 기각하지 못한다.

[프로그램 8.1] [보기 8.2] 젖소용 사료의 우유생산량 독립표본 비교

```
x=c(44, 44, 56, 46, 47, 38, 58, 53, 49, 35, 46, 30, 41)
y=c(35, 47, 55, 29, 40, 39, 32, 41, 42, 57, 51, 39)
t.test(x,y, var.equal=T, alt="greater")
```

[결과 8.1] [프로그램 8.1]의 수행 결과

```
	Two Sample t-test

data:  x and y
t = 0.8676, df = 23, p-value = 0.1973
alternative hypothesis: true difference in means is greater than 0
```

[프로그램 8.2]는 [보기 8.5] 대응표본의 모평균 차 검정에서 두 아웃도어 회사의 등산화 밑창의 마모 정도를 비교하기 위해서 귀무가설과 대립가설 $H_0 : \mu_D = 0$, $H_1 : \mu_D \neq 0$을 검정하기 위한 t.test( ) 함수를 활용한 R - 프로그램이다. 독립표본의 모평균 검정을 위한 [프로

그램 8.1]과 그 형식이 동일하며 다만 대응표본의 모평균 차 검정을 위한 옵션으로 <paired=T>를 사용하고 있다. 양측 대립가설에 대해선 <alt(ernative) = "two.sided">가 필요하지만 이는 디폴트 옵션이다. [결과 8.2]에 대한 해석은 [보기 8.5]를 참고하겠지만 특히, <p-값 = 0.008539>는 유의수준 0.05보다 매우 작아 귀무가설 $H_o : \mu_D = 0$을 기각하게 된다.

[프로그램 8.2] [보기 8.5] 등산화 밑창 마모율 대응표본 비교

```
x=c(14.0, 8.8, 11.2, 14.2, 11.8, 6.4, 9.8, 11.3, 9.3, 13.6)
y=c(13.2, 8.2, 10.9, 14.3, 10.7, 6.6, 9.5, 10.8, 8.8, 13.3)
t.test(x,y,paired=T)
```

[결과 8.2] [프로그램 8.2]의 수행 결과

```
Paired t-test
data:  x and y
t = 3.3489, df = 9, p-value = 0.008539
alternative hypothesis: true difference in means is not equal to 0
```

[프로그램 8.3]은 독립표본 모비율의 차 검정을 위한 [보기 8.7] 남성 노인과 여성 노인의 항체 생성 모비율에 대한 가설 $H_o : p_1 - p_2 = 0$, $H_1 : p_1 - p_2 < 0$을 검정하기 위한 R-프로그램에서 제공하는 prop.test( ) 함수는 카이제곱 통계량에 바탕을 둔 검정을 제공하므로 8.5절의 내용과 보기에서 살펴본 가설 검정을 충실히 따르기 위해서 [프로그램 8.3]은 z.prop 함수를 만들어 사용하고 있다. 특히, [결과 8.3]에는 검정통계량의 관측값 <-1.450804>와 <p−값=$P[z \leq -1.45] = 0.9265828$>이 나타나 있고 이는 유의수준 0.05보다 매우 크므로 귀무가설 $H_o : p_1 - p_2 = 0$을 기각하지 못하게 된다.

[프로그램 8.3] [보기 8.7] 독감 백신 남성과 여성의 항체율 비교

```
z.prop = function(x1,x2,n1,n2){
numerator = (x1/n1) - (x2/n2)
p.common = (x1+x2) / (n1+n2)
denominator = sqrt(p.common * (1-p.common) * (1/n1 + 1/n2))
z.prop.ris = numerator / denominator
return(z.prop.ris)}
n1 <- 113
n2 <- 139
x <- 34
y <- 54
z=z.prop(x, y, n1, n2)
z
pvalue=1-pnorm(z)
pvalue
```

[결과 8.3] [프로그램 8.3]의 수행 결과

```
-1.450804
0.9265828
```

8.7 연습문제

8.1 다음 중 가장 적절한 가설 검정 방법을 선택하여라.

a. 소표본에서 모분산이 동일한 경우의 독립표본
b. 대표본에서 독립표본
c. 소표본에서 모분산이 다른 경우의 독립표본
d. 대응표본
e. 독립표본의 모비율

(1) 한 고등학교의 특성화 교육과정 중 영문에세이 과정의 평균 성적은 남학생보다 여학생이 더 높다고 알려져 있다. 정규분포를 따르며 모분산이 동일하다는 가정 아래 임의로 선발한 남학생 25명의 평균은 76점이고 표준편차는 4점이며, 여학생 29명의 평균은 82점이고 표준편차는 3점이다.

(2) 미국의 숲 100곳 중 56곳이 침엽수로 이루어져 있었고, 멕시코의 숲 80곳 중 40곳이 침엽수로 이루어져 있었다. 미국이 멕시코보다 침엽수가 더 많다고 할 수 있는가?

(3) 10대들은 성인보다 평균적으로 잠을 더 많이 잔다고 알려져 있다. 이를 확인하기 위하여 10대 16명과 성인 12명을 대상으로 연구한 결과, 10대의 평균 수면시간은 8.9시간이고 표준편차는 1.2시간이었으며, 성인의 평균 수면시간은 6.9시간이고 표준편차는 0.5시간이었다.

(4) 다이어트 프로그램의 효과를 알아보기 위하여 49명은 운동요법을, 36명은 식이요법을 시행하였다. 모집단의 표준편차는 각각 2kg과 3kg이다. 운동요법이 식이요법보다 더 많은 체중 감소 효과를 나타내는지를 알고자 한다.

(5) A학원에서 토익 수업을 수강하는 12명의 학생을 대상으로 수강 전 토익 점수와 수강 후 토익 점수를 조사하였다. 토익 수업을 수강한 후, 토익 점수가 증가했는지를 알아보고자 한다.

8.2 새로운 에이즈 예방약의 효과를 알아보기 위하여 실험군인 HIV 양성 환자 224명에게 약을 처방하였더니 그 중 45명의 환자가 4년 후에 에이즈로 발전하였다. 반면에 대조군인 HIV 양성 환자 224명에게는 약을 처방하지 않았으며 그 중 68명이 4년 후에 에이즈로 발전하였다. 이 약이 4년 후에 에이즈로 발전하는 환자의 비율을 줄였는지 아니면 실험군과 대조군의 차이가 없는지를 검정하고자 한다.

(1) 귀무가설 H_0와 대립가설 H_1을 서술하라.

(2) 검정통계량을 구하여라.

(3) 유의수준 5%에서 어떠한 결론을 내릴 수 있는가?

8.3 투수 2명의 투구 속도를 비교하고자 한다. 투수들이 각각 14개의 공을 던진 결과 투수 X의 평균은 86mph이고 표준편차는 3mph이며, 투수 Y의 평균은 91mph이고 표준편차는 7mph이다. 야구팀 감독은 투수 Y의 투구 속도가 투수 X보다 더 빠르다고 믿어도 되는지 알고자 한다.

(1) 귀무가설 H_0와 대립가설 H_1을 서술하라.

(2) 자유도를 구하여라.

(3) 검정통계량을 구하여라.

(4) 유의수준 5%에서 어떠한 결론을 내릴 수 있는가?

8.4 두 음료 중 어떤 것이 설탕이 더 첨가되었는지 확인하기 위하여 모분산이 같고 정규분포를 따르는 모집단으로부터 음료 A 13개와 음료 B 6개를 추출하여 음료의 성분을 분석하였다. 음료 A에 포함된 설탕의 양의 평균은 36g이고 표준편차는 0.6g이며, 음료 B에 포함된 설탕의 양의 평균은 38g이고 표준편차는 0.8g이다. 음료 B가 음료 A보다 설탕이 더 첨가되었다고 할 수 있는지 알고자 한다.

(1) 귀무가설 H_0와 대립가설 H_1을 서술하라.

(2) 검정통계량을 구하여라.

(3) 유의수준 1%에서 어떠한 결론을 내릴 수 있는가?

8.5 미국의 한 의료센터에서 작성한 보고서에 따르면 1900년도에 태어난 백인의 평균 기대 수명은 47.6세이고 흑인의 평균 기대 수명은 33.0세라고 한다. 특정 지역을 선택하여 1900년도에 태어난 백인 124명과 흑인 82명의 수명을 조사하였다. 백인의 평균 수명은 45.3세이고 표준편차는 12.7세이며, 흑인의 평균 수명은 34.1세이고 표준편차는 15.6세로 나타났다. 백인과 흑인의 평균 수명이 같은지를 검정하여라.

(1) 이 검정은 평균에 대한 가설 검정인가, 비율에 대한 가설 검정인가?

(2) 이 검정은 단측 검정인가, 양측 검정인가?

(3) 귀무가설 H_0와 대립가설 H_1을 서술하라.

(4) 이 검정은 어떤 분포를 사용하는가?

(5) 검정통계량을 구하여라.

(6) 유의수준 5%에서 어떠한 결론을 내릴 수 있는가?

8.6 지역 암 센터들은 유방암이 발병한 여성의 수가 2012년보다 2013년에 더 많아졌다고 말한다. 2012년과 2013년을 비교한 결과는 다음과 같고, 이들의 차 D에 대한 평균은 -174.615이고 표준편차는 3.467이다.

지역	2012	2013	지역	2012	2013
서울	15540	15710	제주	3160	3300
부산	7090	7430	세종	4680	5070
울산	3570	3580	창원	3320	3630
대전	6970	7310	전주	2630	2690
대구	6190	6280	청주	2150	2280
인천	15050	14980	포항	1990	2080
광주	3450	3720			

(1) 귀무가설 H_0와 대립가설 H_1을 서술하라.

(2) 검정통계량을 구하여라.

(3) 유의수준 5%에서 어떠한 결론을 내릴 수 있는가?

8.7 마이크로소프트 회사는 연령대별 윈도우 XP 사용의 차이를 알아보기 위해 조사를 실시하였다. 임의로 선택된 조사 참가자들을 다음과 같이 두 연령대 그룹으로 나누었다. 먼저 16세 이상 29세 이하는 628명 중 7%가 윈도우 XP를 사용하고 있었으며, 30세 이상은 2,309명 중 11%가 윈도우 XP를 사용하고 있었다. 두 연령대 그룹의 윈도우 XP 사용은 차이가 있는가?

(1) 귀무가설 H_0와 대립가설 H_1을 서술하라.

(2) 검정통계량을 구하여라.

(3) 유의수준 5%에서 어떠한 결론을 내릴 수 있는가?

8.8 두 금속의 녹는 온도를 비교하기 위하여 정규분포를 따르며 모분산이 동일하다는 가정 아래 두 금속을 각각 15개씩 추출하여 실험하였다. 금속 A의 녹는 온도의 평균은 800°F이고 표준편차는 95°F이며, 금속 B의 녹는 온도의 평균은 900°F이고 표준편차는 105°F이다.

(1) 귀무가설 H_0와 대립가설 H_1을 서술하라.

(2) 검정통계량을 구하여라.

(3) 유의수준 1%에서 어떠한 결론을 내릴 수 있는가?

(4) 유의수준을 5%로 수정하면 결론은 어떻게 달라지는가?

8.9 2년제 대학에 등록한 학생의 수보다 4년제 대학에 등록한 학생의 수가 더 많다고 알려져 있다. 2년제 대학 35곳을 조사한 결과, 평균 등록 학생 수는 5,068명이고 표준편차는 4,777명이다. 4년제 대학 35곳을 조사한 결과, 평균 등록 학생 수는 5,466명이고 표준편차는 8,191명이다. 유의수준 5%에서 검정하여라.

(1) 귀무가설 H_0와 대립가설 H_1을 서술하라.

(2) 검정통계량을 구하여라.

(3) 유의수준 5%에서 어떠한 결론을 내릴 수 있는가?

8.10 맞벌이 부부의 결혼 만족도에 관한 연구의 설문 문항 중 '나는 육아분담에 대하여 만족한다'라는 문항이 있다. 보기는 1은 매우 만족, 2는 만족, 3은 보통, 4는 불만족, 5는 매우 불만족으로 이루어져 있다. 부부 10쌍의 응답 결과가 다음과 같을 때, 남편과 아내의 응답에 차이가 있는지를 검정하여라.

남편	2	2	1	1	2	1	3	1	2	3
아내	4	1	1	4	5	2	1	5	2	5

(1) 귀무가설 H_0와 대립가설 H_1을 서술하라.

(2) 검정통계량을 구하여라.

(3) 유의수준 5%에서 어떠한 결론을 내릴 수 있는가?

(4) 유의수준을 10%로 수정하면 결론은 어떻게 달라지는가?

8.11 새로운 혈압약이 혈압을 낮추는데 효과적인지 알아보고자 6명의 환자를 대상으로 혈압을 측정하고 새로운 약을 복용하게 하였다. 12주 후, 동일한 6명의 환자의 혈압을 다시 측정하였다. 단, 이 연구에서는 수축기 혈압만을 비교하였다.

환자	A	B	C	D	E	F
복용 전	161	162	165	162	166	171
복용 후	158	159	166	160	167	169

(1) 귀무가설 H_0와 대립가설 H_1을 서술하라.

(2) 검정통계량을 구하여라.

(3) 유의수준 1%에서 어떠한 결론을 내릴 수 있는가?

9장

분산분석 : 여러집단의 비교

R과 함께하는

통계학의 이해

9.1 여러 집단의 비교

두 집단 또는 두 처리 간의 비교는 8장의 독립표본과 대응표본의 검정에서 다루었다. 이 장에서는 여러 집단 또는 여러 처리를 비교할 때 적용되는 분산분석(analysis of variance, ANOVA)에 대해 알아보고자 한다. 예를 들어 대기오염이 어린이의 신체 발육에 영향을 미치는 지를 알아보기 위해 공업단지 내의 지역을 대기오염의 정도에 따라 4등급으로 나누고 그곳에 거주하는 어린이의 성장을 비교하여 보는 것이다. 또 다른 예로 스포츠 센터의 4가지 수영강습법(강의 및 비디오, 강의, 비디오, 강의 및 비디오 없이)의 효과를 비교하는 것이다. 이를 경우 여러 처리방법들을 비교할 때 한 번에 두 개씩 짝을 지어 두 처리 간의 비교처럼 하는 것은 올바른 방법이 아니다. 일반적으로 t개의 처리에 의해서 구해진 독립표본에 대하여 처리평균 간 차이의 유의성을 비교 평가하는 방법으로 1920년대 피셔(R.A. Fisher)가 제안한 방법이 분산분석이다. 8장의 두 집단의 모평균 비교를 넘어 세 집단 이상의 t개 집단의 모평균 비교를 위해 다음과 같은 가설을 가정하게 된다.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_t$$

특히, 여러 처리에 의해 독립표본이 잘 얻어지기 위해서는 철저하게 계획되고 확률화(randomization) 원칙을 따르는 실험계획(experimental design)이 이루어져야 한다. 실험계획에는 일원, 이원, 라틴방격, 요인 등 다양하다. 이 장에서는 가장 기초가 되는 일원계획(one-way design)인 일원 분산분석에 대해서 논의하고자 한다.

■ 처리(treatment) 또는 요인(factor)

처리(또는 요인)란, 실험계획에서 대상에 대해 실시할 실험의 조건, 방법 또는 특성을 말한다.

9.2 일원 분산분석

[표 9.1]은 일원 분산분석을 위해 얻어진 t개 처리에 의한 독립표본의 구조이다. 일원 분산분석을 완전확률화계획(completely randomized design)라고도 하는데 이는 t개의 독립표본이 t개의 처리(treatment)에 대한 반응을 나타내는 여러 개 모집단으로부터 독립적인 표본추출에 의해서 얻어졌다고 보기 때문이다.

[표 9.1] 일원 분산분석을 적용할 t개 처리의 독립표본의 구조

처 리	1	2	$\cdots$	t
자 료	y_{11} y_{12} $\vdots$ y_{1n_1}	y_{21} y_{22} $\vdots$ y_{2n_2}	$\cdots$ $\cdots$ $\cdots$	y_{t1} y_{t2} $\vdots$ y_{tn_t}
평 균	$\overline{y}_1$	$\overline{y}_2$		$\overline{y}_t$
총평균	$\overline{y}=\dfrac{\text{관찰값의 총합}}{n_1+n_2+\cdots+n_t}=\dfrac{n_1\overline{y}_1+\cdots+n_t\overline{y}_t}{n}$			

일원 분산분석에서는 전체 $n=n_1+n_2+\cdots+n_t$ 개 실험 또는 조사 대상으로부터 n_1 개가 임의로 표본추출 되어 처리 1에 적용되고, 나머지 남아 있는 $n-n_1$개 대상으로부터 n_2 개가 임의로 추출되어 처리 2에 적용된다. 같은 방법으로 반복하여 최종 남은 n_t 개가 처리 t에서 적용된다. 따라서 연구자는 일원 분산분석에서는 비교하고자 하는 t 개의 처리를 적용할 각각 크기가 n_1, n_2, …, n_t 인 독립표본을 얻게 된다.

[표 9.1]에서 처리 i에 대한 j번째 자료값을 y_{ij} $(i=1,\ldots,t;\ j=1,\ldots,n_i)$로 나타내면 y_{ij}에 대한 식 (9.1)의 수학적 모형을 세울 수 있다.

$$y_{ij} = \mu_i + e_{ij}\,,\ j = 1,\cdots,n_i\,,\ i = 1,\cdots,t \tag{9.1}$$

여기서 μ_i 는 i 번째 처리에 대한 효과(effect)이며 평균 반응값을 나타내고 오차항 e_{ij} 는 서로 독립이며 정규분포 $N(0,\sigma^2)$ 을 따른다.

이젠 t개의 처리 간의 비교는 이들의 t개의 효과 $\mu_i\,,\ i = 1,\cdots,t$ 를 비교하는 것으로 다음과 같은 귀무가설과 대립가설을 검정하는 것이다.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_t\,,\ H_1 : \text{ 모든 } \mu_i \text{가 다 같은 것은 아니다.}$$

이를 위해선 실제로 [표 9.1]의 아래쪽에 요약되어 있는 처리들의 표본평균

$$\overline{y}_1\,,\ \overline{y}_2\,,\ \ldots\,,\ \overline{y}_t$$

이 서로 얼마나 통계적으로 다른 지를 분산분석(ANOVA)를 통하여 살펴보는 것이다.

분산분석을 위한 전반적인 공식은 다음과 같은 기본적인 분해로 출발한다.

$$(y_{ij} - \overline{y}) = (\overline{y}_i - \overline{y}) + (y_{ij} - \overline{y}_i)$$

이 분해의 양변을 제곱하면 다음과 같다.

$$(y_{ij} - \overline{y})^2 = (\overline{y}_i - \overline{y})^2 + (y_{ij} - \overline{y}_i)^2 + 2(\overline{y}_i - \overline{y})(y_{ij} - \overline{y}_i)$$

그런데 $j = 1,\cdots,n_i$ 에 대해 합하면 위 식의 우변 맨 끝항은 $\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i) = 0$ 이므로 다음과 같이 분해된다.

$$\underset{\text{전체제곱합}}{\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2} = \underset{\text{처리제곱합}}{\sum_{i=1}^{t}n_i(\overline{y}_i-\overline{y})^2} + \underset{\text{오차제곱합}}{\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)}$$

전체제곱합(total sum of squares, TSS)는 다음과 같이 정의된다.

$$TSS = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2 \tag{9.2}$$

처리제곱합(sum of squares for the treatment)과 오차제곱합(sum of squares for the error)은

$$SST = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(\overline{y_i}-\overline{y})^2 = \sum_{i=1}^{t}n_i(\overline{y_i}-\overline{y})^2 \tag{9.3}$$

$$SSE = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_i})^2 = TSS - SST \tag{9.4}$$

로 정의된다.

식 (9.3)과 식 (9.4)의 제곱합은 모분산 σ^2으로 나누면 각각 자유도가 $t-1$과 $n-t$인 카이제곱 분포를 따른다고 알려져 있다. 그리고 이들 제곱합을 각각의 자유도로 나눈 값 $MST = SST/(t-1)$와 $MSE = SSE/(n-t)$를 제곱평균(mean square)이라 한다. 특히, 식 (9.4)의 오차제곱합은 σ^2의 추정값으로 사용되며 또한 유의한 차를 나타내기 위해서는 처리제곱평균이 얼마나 커야 하는가에 대한 표준적 척도로 이용된다.

끝으로 귀무가설 $H_0: \mu_1 = \mu_2 = \cdots = \mu_t$ 하에 검정통계량

$$F = \frac{MST}{MSE} = \frac{SST/(t-1)}{SSE/(n-t)} \tag{9.5}$$

는 자유도 $(t-1, n-t)$인 F-분포를 따르며 유의수준 α%에서 귀무가설 H_0: $\mu_1 = \mu_2 = \cdots = \mu_t$ 를 검정을 하기 위한 기각역은 다음과 같다.

$$f \geq F_\alpha(t-1, n-t) \tag{9.6}$$

여기서 f는 식 (9.5)의 검정통계량의 관측값이고 기각치 $F_\alpha(t-1, n-t)$는 분자 자유도 $df_1 = t-1$과 분모 자유도 $df_2 = n-t$인 F-분포의 상위 α%를 나타낸다.

[표 9.2]는 지금까지 설명한 일원 분산분석에 대한 계산을 요약하고 있는 분산분석표이다.

[표 9.2] 일원 분산분석표

요 인 (Source)	자유도 (df)	제곱합 (Sum of Squares)	평균제곱합 (Mean Square)	분산비 (F value)
처 리 (Model)	$t-1$	SST	$MST = SST/(t-1)$	$F = \frac{MST}{MSE}$
오 차 (Error)	$n-t$	SSE	$MSE = SSE/(n-t)$	
전 체 (Total)	$n-1$	TSS		

실제 표본자료를 통하여 [표 9.2]의 분산분석표를 얻기 위해 계산기를 가지고 계산하기란 복잡한 일이다. 이를 9.3절에서 R-프로그램의 anova( ) 함수를 활용하여 [보기 9.1]과 [보기 9.2]의 일원 분산분석을 위한 [프로그램 9.1]과 [프로그램 9.2]가 주어져 있다. 이들의 수행 결과인 [결과 9.1]과 [결과 9.2]에서는 분산분석표를 제공하고 분산비에 의한 식 (9.6)의 기각역에 따른 검정 대신에 검정통계량의 관측값 f의 유의확률인 $P[F \geq f] = p-$값을 제공하고 있다. 따라서 다음의 보기에서도 자세한 계산과정은 생략하고 분산분석표를 이용하여 그 결과를 해석하는 데 초점을 맞추기로 한다.

보기 9.1

네 종류의 가축사료에 의한 병아리의 체중 증가를 알아보려 한다. 임의로 추출된 다섯 마리씩 되어 있는 병아리 4개 그룹에 서로 다른 사료 한 가지만 주어 일정 기간 후 체중 증가량을 측정한 표본자료가 다음과 같이 주어져 있다. 유의수준 0.05에서 사료의 효과에 대한 검정을 시행하라.

사료	A	B	C	D
	55.0	61.0	169.0	42.0
	49.0	112.0	137.0	97.0
	42.0	30.0	169.0	81.0
	21.0	89.0	85.0	95.0
	52.0	63.0	154.0	92.0
평균	43.8	71.0	142.8	81.4

1) 4종류의 사료 A, B, C, D는 처리가 t=4개임을 말하며 사료의 효과가 같다는 다음의 귀무가설을 세울 수 있다.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

2) $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$을 F-검정하기 위한 일원 분산분석표는 다음과 같다.

요인	자유도	제곱합	평균제곱합	분산비
처 리	3	26235	8745	12.105
오 차	16	11559	722.4	
전 체	19	37794		

3) 분산분석표로부터 검정통계량 관측값 f=12.105는 자유도 (3, 16)을 따르며 유의수준 0.05에서 기각치 $F_{0.05}(3, 16)$ = 3.24보다 크므로 귀무가설은 기각되고 사료의 효과에는 차이가 있다는 결론을 짓게 된다.

4) 분산분석의 결과 사료의 효과에 차이가 있으므로 평균을 보면 사료 C에 의한 병아리의 평균 체중 증가가 가장 높음을 알 수 있다.

보기 9.2

스포츠 센터의 수영 강습생중 수준이 비슷한 24명을 임의 추출하여 42 수준으로 나눈 후 네 가지 방법을 적용하여 강습하였다. 강습법에 차이가 있는지, 어느 것이 효과적인지에 관심이 있다. 1개월의 과정이 끝난 후 이들 수료자들에게 25m를 접영으로 수영하게 하여 걸린 시간을 기록한 자료는 다음과 같다.

강습법	시간(초)
강의 및 비디오 촬영	21.4 20.1 21.1 19.6 21.8 19.0
강의	17.8 19.3 19.1 18.8 18.3 19.0
비디오 촬영	18.9 20.3 19.1 19.6 20.0 20.1
강의 및 비디오 촬영 없이	19.9 18.4 18.0 17.9 20.2 19.5

1) 네 가지 수영 강습법 A, B, C, D는 처리가 t=4개임을 말하며 강습법의 효과가 같다는 다음의 귀무가설을 세울 수 있다.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

2) $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$을 F-검정하기 위한 일원 분산분석표는 다음과 같다.

요 인	자유도	제곱합	평균제곱합	분산비
처 리	3	11.42	3.81	5.31
오 차	20	14.35	0.72	
전 체	23	25.77		

3) 분산분석표로부터 검정통계량 관측값 $f=5.31$은 자유도 (3, 20)을 따르며 유의수준 0.05에서 기각치 $F_{0.05}(3, 20)$ = 3.10보다 크므로 기각역을 만족한다.

4) 분산분석의 결과 귀무가설은 기각되고 수영 강습법의 효과에는 차이가 있다고 볼 수 있다.

9.3 R-프로그램 실습

[프로그램 9.1]은 [보기 9.1]의 네 종류의 사료에 의한 병아리 체중 증가의 일원 분산분석에 대한 aov( ) 함수를 활용한 R-프로그램을 보여주고 있다. 결과에 대한 해석은 [보기 9.1]을 참고로 하면 되나 특히, R-프로그램에서는 기각치를 활용한 검정을 제공하는 대신에 $p-$값을 제공하고 있다. [결과 9.1]에서 <Pr(>F)>에 대한 <0.000218>이 $p-$값으로 이는 확률값 $P[F \geqq 12.105]$에 해당한다. 그리고 <***>는 유의수준 $\alpha = 0.0001$보다도 작은 경우를 나타내고 있다. 다른 유의확률에 대한 것의 표시도 결과 아래쪽에 <Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '>에 잘 나타나 있다.

[프로그램 9.1] [보기 9.1] 네 종류의 사료에 의한 병아리 체중 증가의 일원 분산분석

```
wt = c(55, 49, 42, 21, 52, 61, 112, 30, 89, 63, 169, 137, 169, 85, 154, 42, 97,
81, 95, 92)
feed = c(rep("A",5), rep("B",5), rep("C",5), rep("D",5))
chicken = data.frame(wt,feed)
chicken
results = aov(wt ~ feed, data=chicken)
anova(results)
```

[결과 9.1] [프로그램 9.1]의 수행 결과

```
   Analysis of Variance Table

Response: wt
          Df Sum Sq Mean Sq F value   Pr(>F)
feed       3  26235   8745.0  12.105 0.000218 ***
Residuals 16  11559    722.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

[프로그램 9.2]는 [보기 9.2]의 접영에 대한 네 가지 수영 강습법에 의한 1개월 뒤 수영 기록에 대한 일원 분산분석을 위한 R-프로그램으로 [프로그램 9.1]과 유사하다. 다만

<summary(unstack(swim))>에 의해서 네 가지 수영 강습법 A, B, C, D별로 요약 통계량이 [결과 9.2]의 위쪽에 제공되고 있다. 이는 강습법의 효과가 같다는 귀무가설이 기각될 경우 어느 강습법의 효과가 그나마 나은 지에 대한 정보를 제공하게 된다. [결과 9.2]에 따르면 <$p-$값 0.007437>은 $\alpha=0.05$보다 훨씬 작으므로 귀무가설을 기각하게 되며 네 가지 강습법에 대한 요약 통계량 중 줄이 그어진 평균을 보면 실제로 강습법 B에 의한 평균 수영기록 <18.72>이 제일 좋음을 짐작할 수 있다.

[프로그램 9.2] [보기 9.2] 네 가지 수영 강습법에 의한 기록에 대한 일원 분산분석

```
time=c(21.4, 20.1, 21.1, 19.6, 21.8, 19.0, 17.8, 19.3, 19.1, 18.8,
 18.3, 19.0, 18.9, 20.3, 19.1, 19.6, 20.0, 20.1, 19.9, 18.4, 18.0,
 17.9, 20.2, 19.5)
lecture = c(rep("A",6), rep("B",6), rep("C",6), rep("D",6))
swim = data.frame(time,lecture)
summary(unstack(swim))
results = aov(time ~ lecture, data=swim)
anova(results)
```

[결과 9.2] [프로그램 9.2]의 수행 결과

```
     A              B              C              D
 Min.   :19.00   Min.   :17.80   Min.   :18.90   Min.   :17.90
 1st Qu.:19.73   1st Qu.:18.43   1st Qu.:19.23   1st Qu.:18.10
 Median :20.60   Median :18.90   Median :19.80   Median :18.95
 Mean   :20.50   Mean   :18.72   Mean   :19.67   Mean   :18.98
 3rd Qu.:21.32   3rd Qu.:19.07   3rd Qu.:20.07   3rd Qu.:19.80
 Max.   :21.80   Max.   :19.30   Max.   :20.30   Max.   :20.20

Analysis of Variance Table

Response: time
          Df Sum Sq Mean Sq F value   Pr(>F)
lecture    3 11.423  3.8078   5.307 0.007437 **
Residuals 20 14.350  0.7175
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

9.4 연습문제

9.1 3명의 학생에게 영양실험을 위한 실험용 쥐 5마리가 주어졌다. 실험 전 쥐들의 무게를 측정하였으며 단위는 g이다. 영희는 제조법 A, 철수는 제조법 B, 준수는 제조법 C로 만든 음식을 각각 자신의 쥐에게 먹였다. 일정기간이 지난 후, 쥐들의 무게를 다시 측정하였으며 실험 전보다 실제 증가한 무게를 기록한 결과는 아래 표와 같다.

A	B	C
43.5	47.0	51.2
39.4	40.5	40.9
41.3	38.9	37.9
46.0	46.3	45.0
38.2	44.2	48.6

(1) 귀무가설 H_0와 대립가설 H_1을 서술하라.
(2) 분자 자유도와 분모 자유도를 계산하라.
(3) 이 검정은 어떤 분포를 사용하는가?
(4) 처리제곱합(SSt)과 오차제곱합(SSE)을 계산하라.
(5) 검정통계량을 구하여라.
(6) 유의수준 10%에서 어떠한 결론을 내릴 수 있는가?

9.2 다음은 4가지 종류의 잡지의 페이지 수를 기록한 표이다.

(1) 유의수준 5%에서 4가지 종류의 잡지의 평균 페이지 수가 같은지를 검정하여라.

(2) 다른 종류들과 가장 차이가 큰 패션 잡지를 제외하고, 나머지 3가지 종류의 잡지의 평균 페이지 수가 같은지를 유의수준 5%에서 검정하여라.

종류	패션	요리	음악	과학
	172	87	82	104
	286	94	153	136
	163	123	87	98
	205	106	103	207
	197	101	96	146
평균	204.6	102.2	104.2	138.2

9.3 다음은 5개의 농구 팀에서 각 선수들이 얼마나 높이 점프할 수 있는지(단위: 인치)를 조사한 결과이다. 제시된 분산분석표의 빈 칸을 채우시오.

A팀	B팀	C팀	D팀	E팀
36	32	48	38	41
42	35	50	44	39
51	38	39	46	40

분산분석표:

요인	자유도	제곱합	평균제곱합	분산비
처리	㉠	195.6	㉣	㉤
오차	㉡	㉢	23.7	
전체	14	432.9		

9.4 다음은 4개 지역 사람들의 한 달 외식 횟수가 같은지를 알고자 조사한 결과이다. 분산분석표를 작성하여라.

서울	대구	부산	울산
6	4	7	8
8	1	3	3
2	5	5	5
4	2	4	1
6		6	7

9.5 수업방식이 다른 통계학 분반의 기말고사 성적의 평균이 같은지를 알고자 한다. 다음은 각 수업방식별로 몇 분반을 임의로 선택하여 기말고사 성적을 정리한 표이다. 유의수준 5%에서 검정하라.

온라인 수업	혼합 수업 (온라인+면대면)	면대면 수업
72	83	80
84	73	78
77	84	84
80	81	81
81		86
		79
		82

9.6 전국에 있는 여자 고등학생들의 100m 달리기 평균 기록이 같은지를 알고자 한다. 5개 지역의 여자 고등학생들을 임의로 조사한 결과는 다음과 같다. 분산분석표를 작성하고, 유의수준 1%에서 검정하라.

강원도	전라도	충청도	경기도	경상도
19.3	19.9	19.4	19.2	20.1
19.1	19.5	19.5	19.6	20.2
19.4	19.4	19.6	19.5	19.6
19.5	19.2	19.1	19.4	19.8

9.7 3가지 유형의 눈 상태에 따라 스키장의 하루 평균 방문자 수가 같은지를 알아보고자 한다.

(1) 분산분석표를 작성하라.

(2) 유의수준 5%에서 어떠한 결론을 내릴 수 있는가?

(3) 유의수준을 10%로 수정하면 결론은 어떻게 달라지는가?

유형	가루 눈	인공 눈	딱딱한 눈
	1210	2107	2846
	1080	1149	1638
	1537	862	2019
	941	1870	1178
		1528	2233
		1382	
평균	1192	1483	1982.8

9.8 시청자들이 자신이 좋아하는 뉴스 채널을 보는 평균 시간(단위: 분)이 같은지를 알고자 한다. 다음은 이를 조사한 결과를 정리한 표이다. 유의수준 5%에서 검정하라.

KBS	MBC	SBS
45	15	72
12	43	37
18	68	56
38	50	60
23	31	51
35	22	

10장

상관분석과 회귀분석 : 두 변수의 관계

R과 함께하는

통계학의 이해

10.1 상관분석

여러 분야의 통계분석에서 두 개 혹은 그 이상의 측정변수들의 관계가 중요한 경우가 많다. 예를 들자면 공장에서 제품을 생산할 때 생산비용과 생산량의 관계, 농사실험에서 쌀의 수확량과 비료의 투여량의 관계, 기술 작업의 수행에서 기술 습득속도와 훈련의 지속시간의 관계, 공정온도와 제품의 강도의 관계가 있다. 따라서 이들 측정변수 간의 관계에 대한 정도를 나타내는 방법과 서로간의 관계에 대한 함수식의 표현이 가능한 지가 주된 관심이다.

먼저 변수의 관계를 알아보는 데 가장 먼저 할 일은 두 변수 X와 Y에 의해서 n개의 대상을 측정한 자료쌍 (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n)을 이차원 평면상에 점으로 나타낸 그림을 그리는 것이다. 이를 산점도라고 부르며, 이 그림을 통하여 변수들의 관련성을 미리 살펴볼 수 있다.

■ 산점도(散點圖, scatter diagram

두 변수 X와 Y에 의해서 n개의 대상을 측정한 자료쌍 (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n)을 이차원 평면상에 점으로 나타낸 그림.

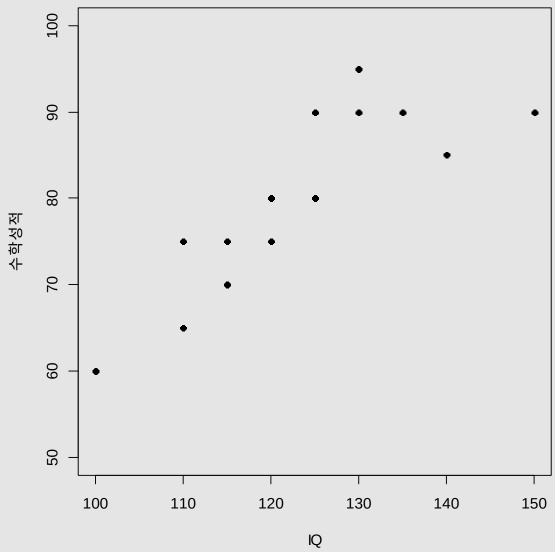
보기 10.1

중학교 1학년 학생들에 대하여 IQ와 수학성적의 관계를 알아보기 위해 15명의 학생을 임의 추출하여 다음의 자료를 얻었다.

학생	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
IQ	110	130	125	120	115	120	125	130	150	140	100	110	115	120	135
수학성적	75	90	80	80	70	75	90	95	90	85	60	65	75	75	90

이에 대한 산점도를 그리시오.

주어진 자료에 대한 산점도는 [그림 10.1]과 같다. 이를 보면 학생들의 IQ가 높을수록 수학성적이 높음을 알 수 있다.



[그림 10.1] IQ와 수학성적의 산점도

일반적으로 산점도에서 자료쌍들이 직선의 형태인 선형에 가까우면 두 변수가 매우 밀접한 관계를 가지고 있다는 증거가 된다. 이와 같이 두 변수 사이의 선형관계를 산점도로부터 대략적으로 알 수 있으나 식 (10.1)과 같이 이를 하나의 수치값인 통계량으로 나타낸 것이 칼 피어슨(Karl Pearson)이 제안한 상관계수(correlation coefficient)이다. 이와 관련된 통계적 분석을 상관분석(correla-tion analysis)이라 한다.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \tag{10.1}$$

식 (10.1)의 통계량 S_{xy}, S_{xx}, S_{yy}는 [보기 10.1]의 자료와 같이 두 변수 X와 Y에 의해서

측정된 n개의 자료쌍이 (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n)일 때 다음과 같이 계산된다. 두 변수의 자료쌍에서 평균을 $\overline{x}$와 $\overline{y}$라 하면 이들로부터 편차들은 $x_1 - \overline{x}, \ldots,\ x_n - \overline{x}$와 $y_1 - \overline{y}, \ldots, y_n - \overline{y}$이고 이들 서로간의 곱의 합은 다음과 같다.

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y} \tag{10.2}$$

그리고 각 변수들의 편차들의 제곱합으로부터

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \tag{10.3}$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 \tag{10.4}$$

를 계산한다. 실제로 식 (10.2), 식 (10.3), 식 (10.4)의 통계량을 $n-1$로 나누게 되면 각각 변수 X와 Y의 (표본)공분산, 변수 X의 (표본)분산 그리고 변수 Y의 (표본)분산이 됨을 알 수 있다.

식 (10.1)의 상관계수 r은 다음의 성질을 가지고 있으며 [그림 10.2]에서 세 종류의 상관에 관한 산점도를 보여주고 있다.

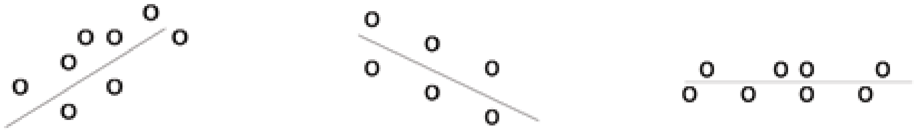
■ 상관계수의 성질

(1) $-1 \leqq r \leqq 1$

(2) $r > 0$: X의 값이 증가함에 따라 Y의 값도 이에 비례하여 증가하는 경향이 있고 이를 양의 상관이라 한다.

(3) $r < 0$: X의 값이 증가함에 따라 Y의 값은 이에 반비례하여 감소하는 경향이 있으며 음의 상관이라 한다.

(4) $r = 0$: 상관이 없는 무상관이다.



(1) $0 < r < 1$　　(2) $-1 < r < 0$　　(3) $r = 0$

[그림 10.2] 산점도: (1) 양의 상관 (2) 음의 상관 (3) 무상관

보기 10.2

[보기 10.1]로부터 중학교 1학년 학생들의 IQ와 수학성적 자료에 대한 상관계수를 다음의 표를 활용하여 구해보라. 여기서 측정 변수로 X=IQ, Y=수학성적이라 하자.

학생	X	Y	X^2	Y^2	XY
1	110	75	12100	5625	825
2	130	90	16900	8100	11700
3	125	80	15625	6400	10000
4	120	80	14400	6400	9600
5	115	70	13225	4900	8050
6	120	75	14400	5625	9000
7	125	90	15625	8100	11250
8	130	95	16900	9025	12350
9	150	90	22500	8100	13500
10	140	85	19600	7225	11900
11	100	60	10000	3600	6000
12	110	65	12100	4225	7150
13	115	75	13225	5625	8625
14	120	75	14400	5625	9000
15	135	90	18225	8100	12150

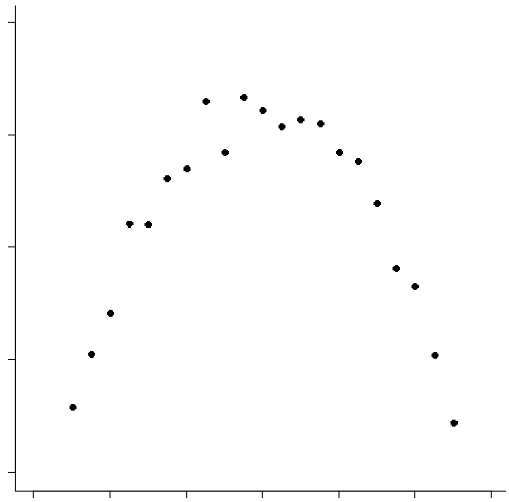
1) $S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$ = 229225 - (1845×1845)/15 = 2290.00

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 96675 - (1195 \times 1195)/15 = 1473.34$$

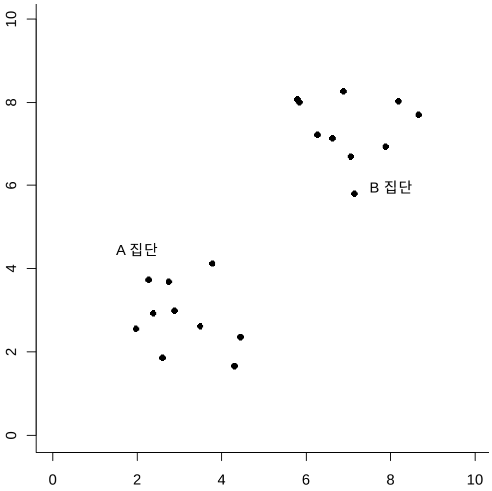
$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = 148525 - (1845 \times 1195)/15 = 1540.00$$

2) 상관계수는 $r = \dfrac{1540}{\sqrt{2290}\sqrt{1473.34}} = 0.8384$가 된다. 이는 두 변수 IQ와 수학성적 간의 높은 양의 상관이 있음을 나타낸다.

상관계수는 두 변수의 관계가 선형에 얼마나 가까운지를 나타내는 수치적 척도이므로 두 변수의 관계가 [그림 10.3]의 곡선 형태이거나, [그림 10.4]처럼 서로 다른 두 집단에 대한 자료쌍이 혼합되어 있는 경우에는 적합하지 못하다. 즉, [그림 10.4]의 전체 자료에 대한 상관계수가 0.897로 매우 높은 양의 상관을 보이므로 상관계수는 산점도와 함께 탐색적으로 사용되어야 한다.



[그림 10.3] 곡선 형태의 산점도



[그림 10.4] 혼합집단의 산점도와 상관계수

특히, [그림 10.4]의 경우 전체 자료의 상관계수는 0.845로 높게 나오지만 집단 A와 B 각각의 상관계수는 0에 가까워 두 집단 내에서의 두 변수는 상관관계가 없는 경우가 된다. 때로는 서로 다른 자료가 잘못 혼합된 경우에도 이와 같은 현상이 발생하므로 반드시 산점도를 살피고 이상 유무를 확인하여야 한다. 또한 원인과 결과의 관계를 생각하지 않고 단순히 두 변수 값들 사이의 상관계수를 구하는 것은 잘못된 결과를 가져오기도 한다. 예를 들어, 어느 지역의 매년 자료를 조사하였더니 범죄율은 계속 높아가고, 껌의 판매량은 줄어들었다면 이 둘은 음의 상관을 보인다고 여겨 범죄율을 낮추기 위해서 껌을 씹도록 권장하게 하는 정책을 수립하는 것은 그럴듯하게 보인다. 반대로 술의 소비량과 성직자들의 수입 간에 상관계수 값이 1에 가깝다고 하여 둘 사이에 높은 상관관계가 있다고 할 수는 없다. 이들 두 변수는 다른 원인에 의하여 나타난 결과들로 두 변수 사이에 실제로는 관계가 그다지 없을 것이다. 이러한 상관관계를 허위상관(spurious correlation)이라고 부른다.

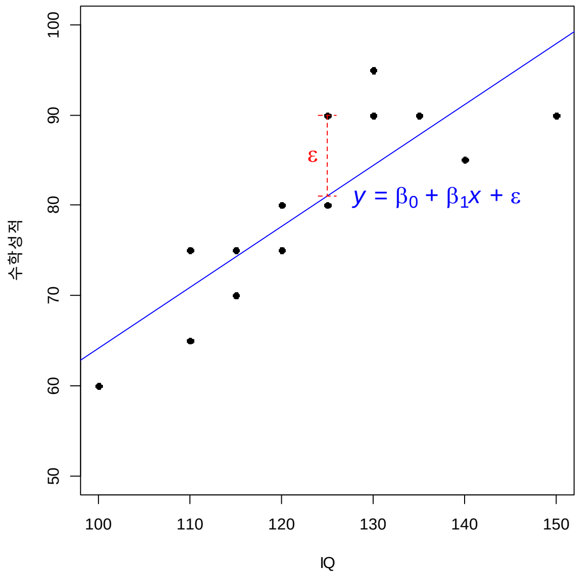
10.2 회귀분석

이 절에서는 10.1절에서 살펴본 두 변수 사이의 선형관계를 함수식으로 나타내는 통계적 방법인 회귀분석을 생각해 보기로 하자.

회귀분석(regression analysis)은 변수들 간의 관계를 나타내는 함수식을 찾아내고 이를 이용하여 예측 및 추론을 하기 위한 분야이다. 회귀(regression)란 말은 영국의 수학자이며 생물학자인 프란시스 갤턴(Francis Galton: 1822 - 1911)이 아버지와 아들의 키에 대한 관계를 연구하면서 처음으로 사용한 용어이다. 갤턴은 키가 매우 큰 부모의 아들은 평균보다는 큰 값을 가지지만 아버지의 키보다는 작은 경향이 있다고 결론 내렸다. 이 결과는 <Regression Toward Mediocrity in Hereditary Stature>라는 제목으로 1885년에 발표되었다. 여기서, <regression toward mediocrity(평범으로 회귀)>란 아들의 키는 평균값에 가까워지려는 경향이 있다는 것을 의미한다. 오늘날 회귀란 단어는 변수들 사이의 관계를 나타내는 뜻으로 사용된다.

회귀분석의 목적을 명확하게 하기 위해 몇 가지 예를 들고자 한다. 아버지의 키와 자녀의 키 사이의 관계식, 또는 IQ와 성적 사이의 관계식을 알아보는 것이다. 이때 아버지의 키는 자녀의 키에 영향을 주게 되므로 설명변수(explanatory variable) 또는 독립변수(independent variable)가 되고 자녀의 키는 영향을 받는 반응변수(response variable) 또는 종속변수(dependent variable)가 된다. 마찬가지로 IQ는 설명변수가 되고 성적은 반응변수가 될 것이다. 일반적으로 설명변수는 x로 반응변수는 y로 나타낸다.

회귀분석에는 단 하나의 설명변수만을 고려한 단순회귀분석(simple regression analysis)과 여러 개의 설명변수들을 고려한 다중회귀분석(multiple regression analysis)이 있다. 이 절에서는 설명변수가 하나이고 설명변수와 반응변수가 선형적인 관계를 가지는 단순회귀만을 다루기로 한다.



[그림 10.5] 산점도와 단순회귀모형식

회귀분석의 첫 단계에서는 [그림 10.5]와 같은 산점도를 그려보는 것이다. 산점도를 그려보면 두 변수 사이의 관계식의 형태를 대략 짐작할 수 있다. 뿐만 아니라 상관계수로 두 변수 사이의 관계가 얼마나 직선에 가까운지를 알아볼 수 있다. 이는 10.1절에서 설명한 바 있다. 두 변수 x와 y의 직선인 선형관계를 나타내는 모형식은 기울기(slope) β_1과 y축에 대한 절편(intercept) β_0로 표현된다. 덧붙여 [그림 10.5]에서 보듯이 모든 자료쌍이 정확히 하나의 직선 위에 있는 것이 아니므로 오차항(error term) ε이 모형식에 추가되어 다음과 같은 모형식을 세울 수 있다.

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{10.5}$$

여기서 x는 설명변수, y를 반응변수라 하며 오차항 ϵ는 평균이 0이고 분산이 σ^2인 표준정규분포 $N(0, \sigma^2)$를 따르는 확률변수이다. 특히, 식 (10.5)의 모형식을 단순회귀모형(simple

regression model)이라고 부르며 이와 같이 오차항을 포함하는 모형을 통계적 모형(statistical model)이라 한다. 단순회귀모형의 기울기 β_1과 절편 β_0는 n개의 자료쌍을 통하여 최소제곱법(method of least square)에 의해서 추정한다. 최소제곱법은 10.3절에서 자세히 설명하도록 한다.

■ 단순회귀모형(simple regression model)

$y = \beta_0 + \beta_1 x + \epsilon$

y: 반응변수(종속변수)

x: 설명변수(독립변수)

ϵ: 오차항으로 평균이 0이고 분산이 σ^2인 정규분포를 따르는 확률변수

두 변수 X와 Y에 의한 n개의 측정된 자료쌍 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$으로부터 최소제곱법에 의해서 구해진 식 (10.5)의 절편과 기울기 β_0와 β_1의 최소제곱추정량은 다음과 같다.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{10.6}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{10.7}$$

여기서 '∧' 표시는 추정량을 나타내며 모자를 쓴 모양이므로 'hat'이라 읽는다. 실제로 식 (10.6)과 식 (10.7)의 추정량의 계산은 상관계수를 구하는 식 (10.1) - 식 (10.4)으로부터 쉽게 이루어진다. 따라서 식 (10.6)과 식 (10.7)의 최소제곱추정량에 의해서 식 (10.5)의 단순회귀모형의 추정 모형식은

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{10.8}$$

이 되며 이를 적합된 회귀식(fitted regression equation)이라 한다.

■ 적합된 회귀식(fitted regression equation)

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}x$$

$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$: 기울기 β_1의 최소제곱추정량

$$\widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x}$$: 절편 β_0의 최소제곱추정량

보기 10.3

[보기 10.1]의 X=IQ와 Y=수학성적에서 IQ가 수학성적에 미치는 영향을 알아보고자 한다. [그림 10.1]의 산점도를 참고하여 모형을 선택한 후 적합된 회귀식을 구하라.

1) [그림 10.1]의 산점도를 보면 두 변수의 관계가 직선 형태에 가까우므로 단순회귀모형 $y = \beta_0 + \beta_1 x + \epsilon$를 선택한다.

2) [보기 10.2]의 상관계수를 구하는 계산과정으로부터 기울기와 절편의 추정치를 구하면

$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} = \frac{1540}{2290} = 0.672,$$

$$\widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x} = \frac{1195}{15} - 0.672 \times \frac{1845}{15} = -3.049$$

3) 적합된 회귀식은

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}x = -3.049 + 0.672x$$

4) IQ가 108인 학생의 수학성적의 추정치는 다음과 같다.

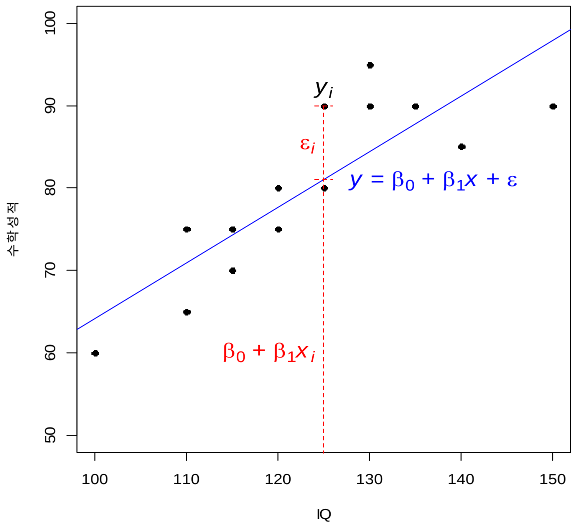
$\hat{y}$ = -3.049 + 0.672×108 = 69.53점

[보기 10.1] - [보기 10.3]의 산점도를 그리거나 상관계수, 적합된 회귀식을 구하는 계산과정을 위한 R-프로그램은 10.5절에 마련해 두었으니 참고하기 바란다.

10.3 최소제곱법과 잔차

단순회귀모형 $y = \beta_0 + \beta_1 x + \epsilon$에서 기울기 β_0와 절편 β_1은 미지의 모수로 n개 자료쌍 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$을 이용하여 이들을 추정하는 문제는 산점도에 직선을 그려 할 수도 있다. 그러나 모든 자료쌍을 동시에 지나는 직선을 구할 수는 없다. 임의의 직선 $y = \beta_0 + \beta_1 x$가 [그림 10.6]에서처럼 산점도에 그려진다고 가정하자. 설명변수 x_i에서 관측된 값 y_i와 이 직선에 의해 예측되는 값 $\beta_0 + \beta_1 x_i$의 차이를 오차 $y_i - \beta_0 - \beta_1 x_i = \epsilon_i$라 한다. 따라서 이 차이를 가장 작게 하는 직선이 이상적인 단순회귀모형의 추정식이 될 것이다. 그러나 모든 자료쌍에서 오차 ϵ_i의 값을 최소화할 수 없기 때문에 총체적으로 작게 하는 방법인 최소제곱법을 이용하게 된다. 우선 임의의 직선 $y = \beta_0 + \beta_1 x$로부터 관찰된 점까지의 총체적인 차이를 나타내는 측도로 다음을 생각할 수 있다.

$$T = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{10.9}$$



[그림 10.6] 각 관측치와 직선 간의 차이

식 (10.9)에서 T의 크기는 임의의 직선 $y=\beta_0+\beta_1 x$를 결정하는 두 모수인 β_0와 β_1에 의존하며 T를 수학적으로 편미분하여 최소화시키는 β_0와 β_1을 결정한다. 이를 $\widehat{\beta}_0$과 $\widehat{\beta}_1$으로 쓰고, 각각을 모수 β_0와 β_1의 *최소제곱추정량*(least squares estimator)이라 하며 실제로 식 (10.6)과 식 (10.7)에 주어져 있다. 이들에 의한 단순회귀 방정식의 추정식인 적합된 회귀식은 식 (10.8)의 $\hat{y}=\widehat{\beta}_0+\widehat{\beta}_1 x$로 주어진다. 10.2절의 [보기 10.3]에서 이미 적합된 회귀식을 구하는 계산의 예를 살펴보았다.

이미 설명했듯이 적합된 회귀선은 $\sum(y_i-\beta_0-\beta_1 x_i)^2$이 최소화된다는 의미에서 자료에 가장 잘 적합된다. 적합된 관측치 $\widehat{y_i}=\widehat{\beta}_0+\widehat{\beta}_1 x_i$와 관측치 y_i의 편차를 잔차(residual)라 부르고 $e_i=y_i-\widehat{\beta}_0-\widehat{\beta}_1 x_i,\ \ i=1,\ldots,n$로 나타내자. 잔차의 성질은 양수일 수도 있고, 음수일 수도 있는데 그 합인 $\sum e_i$은 언제나 0이다.

잔차제곱합(residual sum of squares) $\sum e_i^2$은 10.2절에서 정의한 단순회귀모형의 오차항 ϵ의 분포 $N(0,\sigma^2)$에서 모르는 분산 σ^2의 추정치를 제공한다. 잔차제곱합은 오차에 기인한 제곱합이라 하여 오차제곱합(sum of squares due to error)이라고도 부르며 다음과 같다.

$$SSE=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(y_i-\widehat{y_i})^2 \tag{10.10}$$

다음으로 식 (10.10)으로부터 σ^2의 추정량인 $\hat{\sigma}^2$을 평균제곱오차(mean squared error) MSE라 하며 다음과 같이 정의된다.

$$\hat{\sigma}^2=\frac{SSE}{n-2}=MSE \tag{10.11}$$

참고로 식 (10.10)의 잔차제곱합 SSE의 편리한 계산공식은 식 (10.1)의 상관계수를 구할 때 이용한 통계량 S_{xx}, S_{yy}, S_{xy}에 의해 다음과 같이 요약된다.

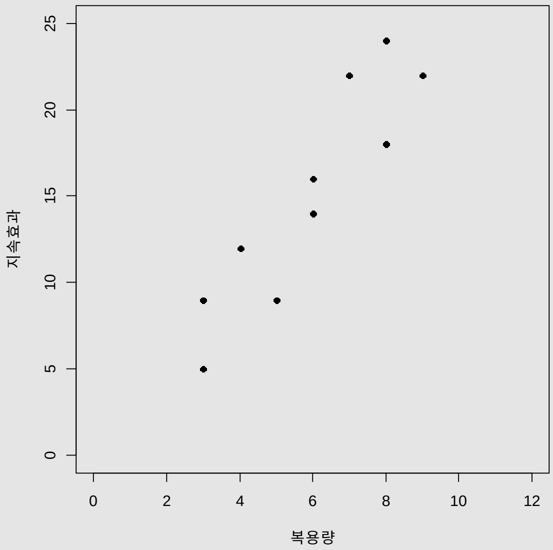
$$SSE=S_{yy}-\frac{S_{xy}^2}{S_{xx}}$$

보기 10.4

새로운 약품을 개발하는 단계에서 약의 지속효과에 영향을 주는 약의 복용량이 어떻게 다른지 알고 싶다. 10명의 환자를 대상으로 실험한 약의 복용량(x)과 약의 지속효과 기간(y)에 대한 자료가 다음과 같이 주어져 있다. 단순회귀모형을 가정하고 최소제곱법을 적용한 적합된 회귀식과 잔차제곱합, 잔차의 분산을 추정하여 보자.

환 자	1	2	3	4	5	6	7	8	9	10
약의 복용량 x	3	3	4	5	6	6	7	8	8	9
약의 지속효과 y	9	5	12	9	14	16	22	18	24	22

1) 약의 복용량(x)과 약의 지속효과 기간(y)에 대한 산점도에 따르면 약의 지속효과 기간 y가 약의 복용량 x에 따라 대체로 증가하는 것으로 보여 단순회귀모형 $y = \beta_0 + \beta_1 x + \varepsilon$을 가정할 수 있다.



2) 다음의 절차에 따라 적합된 회귀식을 계산할 수 있다.

환자	x	y	x^2	y^2	xy	$\hat{y}$	$e=y-\hat{y}$
1	3	9	9	81	27	7.15	1.85
2	3	5	9	25	15	7.15	-2.15
3	4	12	16	144	48	9.89	2.11
4	5	9	25	81	45	12.63	-3.63
5	6	14	36	196	84	15.37	-1.37
6	6	16	36	256	96	15.37	0.63
7	7	22	49	484	154	18.11	3.89
8	8	18	64	324	144	20.85	-2.85
9	8	24	64	576	192	20.85	3.15
10	9	22	81	484	198	23.59	-1.59
합계	59	151	389	2651	1003		0.04

통계량인 $\overline{x}$, $\overline{y}$, S_{xx}, S_{yy}, 그리고 S_{xy}를 계산한다.

$$\overline{x} = 5.9,\ \overline{y} = 15.1$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 = 389 - 10(5.9)^2 = 40.9$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 = 2651 - 10(15.1)^2 = 370.9$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y} = 1003 - 10(5.9 \times 15.1) = 112.1$$

3) β_0와 β_1의 최소제곱추정량과 적합된 회귀식을 구한다.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{112.1}{40.9} = 2.74,$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 15.1 - 2.74 \times 5.9 = -1.07,$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -1.07 + 2.74x$$

적합된 회귀식으로부터 약의 복용량 x의 값에 따른 추정된 약의 지속효과 시간 $\hat{y}$의 값이 절차 2)의 표에 계산되어 있다.

4) 잔차와 잔차제곱합을 계산한다.

잔차 $e_i = y_i - \hat{y}_i = y_i + 1.07 - 2.74x_i$는 절차 2)의 표 마지막 열에 계산되어 있고 이들 잔차제곱합은

$\sum e_i^2 = (1.85)^2 + (-2.15)^2 + (2.11)^2 + \cdots + (-1.59)^2 = 63.653$이다. 이는 다음의 계산과 일치함을 알 수 있다.

$$SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 370.9 - \frac{(112.1)^2}{40.9} = 63.6528$$

5) 잔차의 분산 σ^2 추정은 다음의 평균제곱오차이다.

$$MSE = \frac{SSE}{n-2} = \frac{63.6528}{8} = 7.96$$

10.4 적합된 회귀식의 타당성

가정한 단순회귀모형 $y = \beta_0 + \beta_1 x + \varepsilon$을 최소제곱법에 의해서 추정하여 얻어진 적합된 회귀선 $\hat{y_i} = \hat{\beta_0} + \hat{\beta_1} x_i$의 타당성을 위해 우리는 반응변수에서의 얼마나 많은 변동이 적합된 회귀선에 의해 설명되는지를 제공하는 척도가 요구된다. 특히 반응변수의 관측치 y_i는 다음과 같이 두 요소의 구성으로 볼 수 있다.

y_i	$=$	$(\hat{\beta_0} + \hat{\beta_1} x_i)$	$+$	$(y_i - \hat{\beta_0} - \hat{\beta_1} x_i)$
관측된 y		선형관계에 의해 설명되는 부분		선형관계로부터의 잔차

모든 자료쌍들이 정확하게 직선 위에 놓이는 이상적인 상황이라면 잔차 $y_i - \hat{\beta_0} - \hat{\beta_1} x_i$는 모두 0이다. 이는 y가 완전하게 x에 선형종속 되는 것이며 잔차제곱합인

$$SSE = \sum (y_i - \hat{\beta_0} - \hat{\beta_1} x_i)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}} \tag{10.12}$$

는 선형성으로부터의 불일치의 전체 척도로 생각할 수 있다. y의 총변동은 SSE의 일부가 되는 제곱합

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

에 반영된다. 이를 총제곱합(total sum of squares)이라 하며 SST로 표기한다. SST와

SSE의 차이는

$$S_{yy} - SSE = S_{yy} - \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = \frac{S_{xy}^2}{S_{xx}}$$

으로 주어지며 이는 총변동 중 회귀식에 의해서 설명된 변동으로 회귀제곱합(sum of squares due to regression)이라 하며 SSR로 표기한다. 따라서 y의 총변동 SST는 다음과 같이 분해된다.

$$\begin{array}{ccccccc} & S_{yy} & = & \dfrac{S_{xy}^2}{S_{xx}} & + & SSE & \\ \Leftrightarrow & SST & = & SSR & + & SSE & (10.13) \\ & \text{총제곱합} & & \text{회귀제곱합} & & \text{잔차제곱합} & \end{array}$$

식 (10.3)의 분해를 보면 모든 자료쌍들이 추정된 회귀식 위에 놓인 이상적인 상황에서는 SSE는 0이다. 따라서 총변동 SST 중에서 SSE가 차지하는 변동이 작거나 반대로 SSR이 차지하는 변동이 크면 적합된 회귀식의 타당도는 높다고 할 수 있다. 그러므로 적합된 회귀식의 타당도는 총변동 SST 중에서 회귀식에 의해서 설명되는 변동 SSR이 차지하는 비

$$R^2 = \frac{SSR}{SST} = \frac{S_{xy}^2 / S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx} S_{yy}} \tag{10.14}$$

에 의해서 측정할 수 있으며 이를 결정계수(coefficient of determination)라 한다. 그리고 식 (10.14)의 결정계수는 식 (10.1)에서 정의한 상관계수의 제곱 r^2과 일치함을 알 수 있다. 실제 상관계수 값의 범위가 $-1 \leq r \leq 1$이므로 결정계수의 값은 항상 0과 1 사이에 있게 된다. 결정계수의 값이 1에 가까울수록 자료쌍들이 적합된 회귀식 주위에 밀집해 있음을 의미하며 회귀식이 잘 추정되었고 타당하다고 할 수 있다.

■ 결정계수 R^2 와 상관계수 r 의 관계

$$R^2 = \frac{SSR}{SST} = = \frac{S_{xy}^2}{S_{xx} S_{yy}} = r^2$$

보기 10.5

[보기 10.4]의 10명의 환자에 대한 약의 복용량(x)과 지속효과 시간(y)에 대한 자료의 상관계수를 구하여 적합된 회귀선 $\hat{y} = -1.07 + 2.74x$의 타당도를 구하여 보라.

1) 세 가지 제곱합 $S_{xx} = 40.9$, $S_{yy} = 370.9$, $S_{xy} = 112.1$으로부터 상관계수는

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{112.1}{\sqrt{40.9}\sqrt{370.9}} = 0.91$$

이다.

2) 결정계수는 상관계수의 제곱인 $r^2 = 0.91^2 = 0.83$이다. 이는 10명의 약의 지속효과 기간 y가 갖는 총변동에서 83%를 약의 복용량 x를 사용한 단순회귀모형으로 설명할 수 있음을 나타낸다.

결정계수 외에 적합된 회귀식의 타당도를 설명하는 또 다른 측도로 제곱합의 분해를 이용한 분산분석(ANOVA: analysis of variance)표를 생각할 수 있다.

[표 10.1] 단순회귀모형의 타당성 검정을 위한 분산분석표

요인 (Source)	자유도 (df)	제곱합 (SS)	평균제곱 (MS)	F 값 (F value)
모형 (Model)	1	SSR	$MSR = SSR$	$F = MSR/MSE$
잔차 (Error)	$n-2$	SSE	$MSE = SSE/(n-2)$	
전체 (Total)	$n-1$	SST		

[표 10.1]의 분산분석표에서 $F = MSR/MSE$는 적합된 회귀직선에 의해서 설명되는 부분 SSR과 잔차에 의해서 설명되는 부분 SSE과 비교하는 값이다. 이 값이 크면 적합된 회귀선이 유의함을 나타내며 통계적으로 다음의 귀무가설을 검정하면 된다.

$$H_0 : \beta_1 = 0$$

즉, 단순회귀모형 $y = \beta_0 + \beta_1 x + \epsilon$에서 설명변수 x의 크기에 따라 반응변수 y가 변하는지 변하지 않는지에 관심이 있다. 이는 기울기 β_1이 0인지 아닌지를 알아보는 것과 관련이 있다. 만약에 $\beta_1 = 0$라면 설명변수 x는 반응변수 y를 설명하는 데 의미가 없다. 즉, 단순회귀모형이 적합하지 못함을 의미한다. 단순회귀모형에서는 기울기 $\beta_1 = 0$를 검정하는 것과 모형이 적합한지를 검정하는 것이 동일하다는 점이다.

실제 [표 10.1]의 분산분석표에서 $F = MSR/MSE$의 관측값 f이 유의수준 α에서 기각치 $F_\alpha(1, n-2)$에 의한 다음의 기각역을 만족하면 귀무가설 $H_0 : \beta_1 = 0$를 기각하게 된다.

$$f > F_\alpha(1, n-2)$$

이는 가정된 단순회귀모형이 적합하고 따라서 적합된 회귀식이 유의함을 나타낸다. 10.5절의 R-프로그램에는 summary(lsm)과 anova( ) 함수로부터 검정통계량의 관측값의 유의확

률인 $p-$값$=P[F\geq f]$을 제공되어 이 값이 유의수준 α보다 작아도 앞의 결론을 얻게 된다.

■ 단순회귀모형의 타당성 검정

$y = \beta_0 + \beta_1 x + \epsilon$에서 $H_0 : \beta_1 = 0$의 검정 :

$f > F_\alpha(1, n-2) \Rightarrow$ 귀무가설을 기각하여 단순회귀모형이 적합함을 나타낸다.

여기서, f는 $F = MSR/MSE$의 관측값이다.

보기 10.6

[보기 10.5]의 10명의 환자에 대한 약의 복용량(x)과 지속효과(y)에 대한 자료의 적합된 회귀식 $\hat{y} = -1.07 + 2.74x$의 유의성 검정을 위한 분산분석표를 작성하고 해석하라.

1) [보기 10.5]의 결정계수 계산에서 $S_{xx} = 40.9$, $S_{yy} = 370.9$, $S_{xy} = 112.1$이므로 총제곱합과 회귀제곱합은 다음과 같이 계산된다.

$$SST = S_{yy} = 370.90,\ SSR = \frac{S_{xy}^2}{S_{xx}} = \frac{(112.1)^2}{40.9} = 307.25$$

이로부터 오차제곱합은 변동의 분해공식 $SST = SSR + SSE$에 의해서

$$SSE = 370.90 - 307.25 = 63.65$$

임을 알 수 있다.

2) 단순회귀모형의 타당성 검정을 위한 분산분석표

요인 (Source)	자유도 (df)	제곱합 (SS)	평균제곱 (MS)	F 값 (F value)
모형 (Model)	1	307.25	307.25	38.62
잔차 (Error)	8	63.65	7.96	
전체 (Total)	9	970.90		

3) 분산분석표의 F값의 관측값 38.62이 유의수준 0.05에서 기각치 $F_{0.05}(1,8)=5.32$에 의한 기각역을 만족하므로 귀무가설 $H_0 : \beta_1 = 0$를 기각하게 된다. 따라서 적합된 회귀직선은 유의하다.

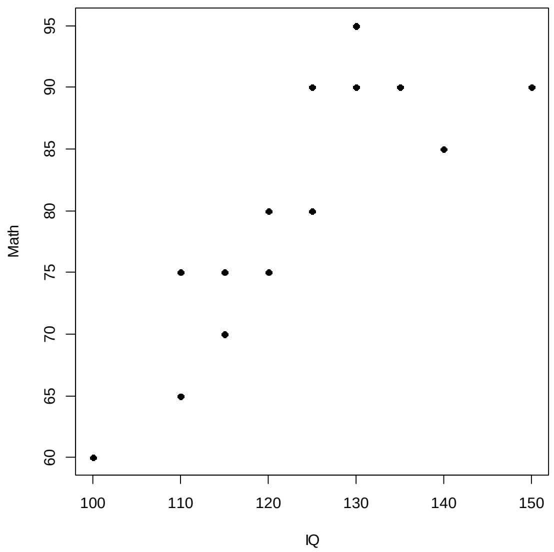
10.5 R-프로그램 실습

[프로그램 10.1]은 IQ와 수학성적에 대한 [보기10.1]과 [보기 10.2]의 산점도와 상관계수를 제공하는 R-프로그램이다. 산점도는 plot( ) 함수를 상관계수는 cor( ) 함수를 사용한다. 그 수행 결과는 [결과 10.1]에 주어져 있다.

[프로그램 10.1] [보기 10.1]-[보기 10.2]의 IQ와 수학성적에 대한 산점도와 상관계수

```
IQ=c(110, 130, 125, 120, 115, 120, 125, 130, 150, 140, 100, 110, 115, 120,
135)
Math=c(75, 90, 80, 80, 70, 75, 90, 95, 90, 85, 60, 65, 75, 75, 90)
plot(IQ, Math, pch=16)
cor(IQ, Math)
```

[결과 10.1] [프로그램 10.1]의 수행 결과




```
0.8384023
```

[프로그램 10.2]는 IQ와 수학성적의 적합된 회귀식을 위한 β_0와 β_1의 최소제곱추정량을 위한 lm( ) 함수와 적합된 회귀선을 원래 산점도에 같이 나타내기 위해 abline( ) 함수를 이용하는 R-프로그램이다. [결과 10.2]의 1) 산점도에 그 결과가 나타나 있다. [결과 10.2]의 2) 적합된 회귀모형의 <Coefficients>에서 실제로 적합된 회귀식을 위한 β_0와 β_1의 최소제곱추정량을 제공한다. (Intercept)와 x에서 각각 $\widehat{\beta_0} = -3.0495$와 $\widehat{\beta_1} = 0.6725$을 얻게 되며

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}x = -3.0495 + 0.6725x$$

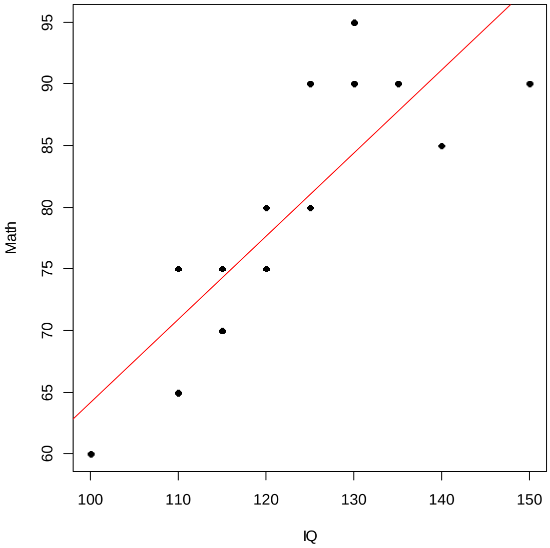
이 구하고자 하는 적합된 회귀식이다. 특히, 모형의 적합성을 위한 귀무가설 $H_0 : \beta_1 = 0$를 검정하기 위한 검정통계량이 [결과 10.2]의 맨 아래 <F-statistic: 30.76 on 1 and 13 DF, p-value: 9.447e-05>에 주어져 있다. $p-$값=0.000094447로 유의수준 5%보다 매우 작아 귀무가설을 기각하기에 충분하며 IQ와 수학성적간의 단순회귀모형이 적합함을 보여준다.

[프로그램 10.2] [보기 10.3]의 단순회귀분석

```
IQ=c(110, 130, 125, 120, 115, 120, 125, 130, 150, 140, 100, 110, 115, 120,
135)
Math=c(75, 90, 80, 80, 70, 75, 90, 95, 90, 85, 60, 65, 75, 75, 90)
x=IQ
y=Math
lsm=lm(y~x)
summary(lsm)
plot(IQ, Math, col=" blue", pch=16, main="Scatter Diagram and Fitted
Regression Line ")
abline(coef(lsm), col="red")
```

[결과 10.2] [프로그램 10.2]의 수행 결과

1) 산점도



2) 적합된 회귀식

```
Call:
lm(formula = y ~ x)

Coefficients:
            Estimate  Std. Error  t value Pr(>|t|)
(Intercept)  -3.0495    14.9894   -0.203    0.842
x             0.6725     0.1213    5.546 9.45e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.803 on 13 degrees of freedom
Multiple R-squared:  0.7029,    Adjusted R-squared:  0.6801
F-statistic: 30.76 on 1 and 13 DF,  p-value: 9.447e-05
```

[프로그램 10.3]은 약의 복용량(x)와 약의 지속효과 기간(y)의 산점도와 단순회귀모형을 적합하는 [보기 10.4]와 모형의 적합성을 보여주는 결정계수를 얻기 위한 [보기 10.5]를 위한 R-프로그램이다. [결과 10.3]의 <Coefficients>의 (Intercept)와 x에서 각각 추정치가 $\hat{\beta}_0 = -1.0709$와 $\hat{\beta}_1 = 2.7408$이므로 따라서 적합된 회귀식은 $\hat{y} = -1.0709 + 2.7408x$이 된다. 다음으로 [결과 10.3]의 <Residual standard error>에서 잔차의 분산 σ^2 추정량인 평균제곱오차

$$MSE = \frac{SSE}{n-2} = \frac{63.6528}{8} = 7.96$$

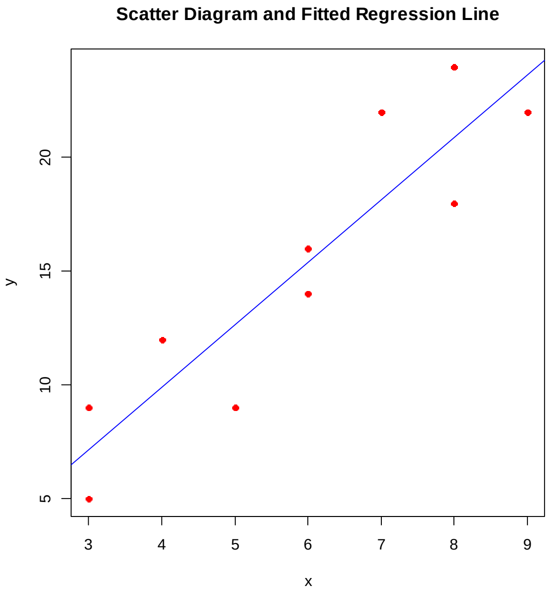
의 $\sqrt{MSE} = \sqrt{7.96} = 2.821$를 얻을 수 있다. 끝으로 <Multiple R-squared>에서 0.8284은 적합된 회귀모형의 타당성을 나타내는 척도인 결정계수 R^2를 나타낸다. 즉, 적합된 회귀식이 총변동에서 82.84%를 설명한다고 보게 된다. 더군다나, 모형의 적합성을 위한 귀무가설 $H_0 : \beta_1 = 0$를 검정하기 위한 검정통계량이 [결과 10.3]의 맨 아래 <F-statistic: 38.62 on 1 and 8 DF, p-value: 0.0002555>에 주어져 있다. $p-$값=0.0002555이 유의수준 5%보다 매우 작아 귀무가설을 기각하며 약의 복용량(x)와 약의 지속효과 기간(y)의 단순회귀모형이 적합함을 보여준다.

[프로그램 10.3] [보기 10.4]-[보기 10.5] 약의 복용량과 약의 지속효과 기간의 단순회귀분석, 결정계수 및 모형의 타당성 검정

```
x=c(3, 3, 4, 5, 6, 6, 7, 8, 8, 9)
y=c(9, 5, 12, 9, 14, 16, 22, 18, 24, 22)
plot(x, y, col="blue", pch=16)
lsm=lm(y~x)
summary(lsm)
plot(x, y, col="red", pch=16, main="Scatter Diagram and Fitted Regression Line")
abline(coef(lsm), col="blue")
```

[결과 10.3] [프로그램 10.3]의 수행 결과

1) 산점도



2)

```
Call:
lm(formula = y ~ x)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.0709     2.7509  -0.389 0.707219
x             2.7408     0.4411   6.214 0.000255 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.821 on 8 degrees of freedom
Multiple R-squared:  0.8284,    Adjusted R-squared:  0.8069
F-statistic: 38.62 on 1 and 8 DF,  p-value: 0.0002555
```

[프로그램 10.4]는 약의 복용량(x)와 약의 지속효과 기간(y)의 회귀분석모형의 타당성을 검정하기 위한 분산분석표를 얻기 위해 anova( ) 함수를 활용하는 R-프로그램이다. 이는 [프로그램 10.3]을 활용해도 충분하며 [결과 10.4]에 대한 해석은 [결과 10.3]을 참고하기 바란다.

[프로그램 10.4] [보기 10.4] 약의 복용량과 약의 지속효과 기간의 회귀분석모형에 대한 타당성

```
x = c(3, 3, 4, 5, 6, 6, 7, 8, 8, 9)
y = c(9, 5, 12, 9, 14, 16, 22, 18, 24, 22)
lsm = lm(y ~ x)
anova(lsm)
```

[결과 10.4] [프로그램 10.4]의 수행 결과

```
1) Analysis of Variance Table

Response: y
          Df Sum Sq  Mean Sq  F value    Pr(>F)
x          1 307.247  307.247  38.615 0.0002555 ***
Residuals  8  63.653    7.957
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

10.6 연습문제

10.1 임의로 선택된 10명의 국가대표 선수들을 대상으로 자신이 받은 후원금의 횟수와 액수(단위:억)의 관계를 알아보고자 다음의 자료를 얻었다.

선수	1	2	3	4	5	6	7	8	9	10
횟수	1	3	2	1	5	5	4	3	1	4
액수	2	8	7	4	13	12	9	9	3	10

(1) 산점도를 그려라.

(2) 상관계수를 구하라.

(3) (2)에서 구한 상관계수로 볼 때, 후원금의 횟수와 액수는 어떠한 관계가 있는가?

10.2 연습문제 10.1에서 후원금의 횟수에 따른 후원금의 액수를 알아보기 위한 회귀식을 구하고, 적합된 회귀식의 결정계수를 구하라.

10.3 다음은 특정 연도에 태어난 사람들의 기대수명을 나타낸 표이다. 태어난 연도에 따른 기대수명을 알아보고자 한다.

태어난 연도	기대수명
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75
1992	75.7
2010	78.7

이때 태어난 연도를 X, 기대수명을 Y라 하면 $\sum_{i=1}^{9} x_i = 34929611$, $\sum_{i=1}^{9} y_i = 45504.02$, $\sum_{i=1}^{9} x_i y_i = 1257635$로 계산되어 진다.

(1) 회귀식을 구하라.

(2) 적합된 회귀식으로부터 2014년에 태어난 사람의 기대수명을 예측하라.

10.4 한 잡지에 음식점 9곳의 할인 쿠폰이 실려 있다. 페이지에 따른 할인 가격(단위:천 원)을 알아보고자 한다.

페이지	4	14	25	32	43	57	72	85	90
할인가격	16	19	15	17	19	15	16	15	17

(1) 산점도를 그려라.

(2) 상관계수를 구하라. 어떠한 관계가 있는가?

(3) 회귀식을 구하라.

(4) 잔차제곱합, 회귀제곱합, 총제곱합을 구하라.

(5) 결정계수를 구하라.

(6) 분산분석표를 작성하고, (3)에서 적합된 회귀식을 검정하라.

10.5 한 전자상가에서 TV 광고를 한다. TV의 사이즈(단위:인치)별 판매 가격(단위:천 원)은 다음의 표와 같다.

사이즈	9	20	27	31	35	40	60
가격	147	197	297	447	1177	2177	2497

(1) TV의 사이즈에 따른 판매 가격을 알아보기 위한 회귀식을 구하라.

(2) (1)에서 적합된 회귀식을 검정하라.

(3) 50인치 TV의 판매 가격을 예측하라.

10.6 다음은 1990년에 미국 남성들의 키(단위:cm)를 측정한 자료이다.

나이	0	2	3	5	7	10	14
키	50.8	83.8	91.4	106.6	119.3	137.1	157.5

(1) 미국 남성들의 만 나이에 따른 키를 알아보기 위한 회귀식을 구하라.
(2) 만 나이가 1세와 11세일 때 키를 예측하라.
(3) 만 나이가 62세일 때 키를 예측하라. 이 결과가 타당한가?

10.7 다음은 연간 10만 명당 사망 운전자 수를 연령별로 정리한 표이다.

연령	17.5	22	29.5	44.5	64.5	80
사망자 수	38	36	24	20	18	28

(1) 연령에 따른 사망 운전자 수를 알아보기 위한 회귀식을 구하라.
(2) 상관계수와 결정계수를 구하라.

10.8 자동차의 무게와 연비의 관계를 알아보기 위해 자동차 12대를 대상으로 무게(단위:파운드)와 연비(단위:mi/g)를 측정한 표이다. 산점도를 그리고, 상관계수를 구하라.

무게	연비
2715	24
2570	28
2610	29
2750	38
3000	25
3410	22
3640	20
3700	26
3880	21
3900	18
4060	18
4710	15

10.9 연습문제 10.8의 자료를 이용하여 다음 물음에 답하라.

(1) 자동차의 무게에 따른 연비를 알아보기 위한 회귀식을 구하라.

(2) (1)에서 적합된 회귀식의 결정계수를 구하라.

(3) 분산분석표를 작성하고, (1)에서 적합된 회귀식을 검정하라

11장

분할표 자료분석 : 범주들의 관계

R과 함께하는

통계학의 이해

11.1 분할표

1.3절의 자료의 이해에서 자료는 측정변수의 성질에 따라 일반적으로 질적 자료와 양적 자료로 나눌 수 있고 특히, 질적 자료는 [표 1.1]에서와 같이 달리기(1=달리기 한 사람, 2=달리기를 하지 않은 사람), 흡연(1=흡연자, 2=비흡연자), 성별(1=남자, 2=여자), 운동량(1=적음, 2=보통, 3=많음)과 같이 몇 개의 범주를 가진 변수에 의해서 측정된 것이며 질적이란 개체인 측정대상이 어느 범주에 들어가는지를 나타내는 것을 의미한다. 일반적으로 여러 사람이나 제품을 대상으로 이루어진 통계조사에서 다음과 같은 몇 가지 경우를 생각할 수 있겠다.

1) 사람들의 종교를 가톨릭, 신교도, 불교, 기타로 분류하는 경우
2) 직업에 대한 만족도를 매우 만족, 보통 만족, 불만족으로 관찰한 경우
3) 공장에서 제조되는 냉장고를 그 성능에 따라 매우 양호, 양호, 약간 결함, 불량으로 관한 경우
4) 교육수준에 따른 당의 선호도를 관찰한 경우

여기서 종교, 만족도, 성능, 교육수준, 선호도는 대상이 어느 범주에 속하는지를 분류하는 질적 변수이며 이를 범주형 변수라 한다. 몇 개의 범주를 가진 이들 범주형 변수에 의해서 관찰된 대상을 분류하여 보면 각 범주에 따른 도수를 가지는 표(table)를 얻게 된다. 다음의 [보기 11.1]과 [보기 11.2]를 통하여 범주형 변수에 의해서 이루어진 표를 살펴보자.

보기 11.2

두 가지 식이요법 A와 B의 효과를 비교하기 위해 150명의 환자를 대상으로 임의 추출된 80명에게는 식이요법 A를 나머지 70명에게는 식이요법 B를 적용하였다. 얼마간의 시간이 흐른 뒤에 식이요법과 각 환자의 건강상태인 2개의 범주형 변수에 따라 자료를 표로 정리하였다. 식이요법은 범주 A와 B로, 환자의 건강상태는 양호, 보통, 불량으로 세 범주로 구성되어 있다.

식이요법	환자의 건강상태			합 계
	양 호	보 통	불 량	
A	37	24	19	80
B	17	33	20	70
합 계	54	57	39	150

[보기 11.1]과 [보기 11.2]에서와 같이 관찰된 자료가 범주형 변수에 따라 표로 정리되어 있을 때 이를 분할표(contingency table)라 부른다. 특히, 2개의 범주형 변수에 의한 분할표를 2원(two-way)분할표라 하며 이외에도 3원분할표 이상의 다원(multi-way)분할표도 있다. 그리고 분할표에서 행에 있는 범주를 행범주, 열에 있는 범주를 열범주라 한다. 만약에 분할표의 행과 열의 수가 각각 r과 c일 때 이러한 표를 $r \times c$ 분할표라 한다. 따라서 이들 분할표는 $r \times c$개의 칸(cell)을 갖게 된다. [보기 11.1]의 표는 3×2 분할표이고, [보기 11.2]의 표는 2×3 분할표로 둘 다 6칸을 가지고 있다.

■ 분할표(contingency table)

관찰된 자료가 범주형 변수에 따라 표(table)로 정리된 것이다. 특히, 2개의 범주형 변수에 의한 분할표를 2원(two-way)분할표라 하며 이외에도 3원분할표 이상의 다원(multi-way)분할표도 있다.

11.2 카이제곱 통계량

[보기 11.1]과 [보기 11.2]의 분할표를 통하여 알고자 하는 점을 요약하자면 다음과 같다.

① 나이에 따라 약의 종류에 따른 선호도가 관계가 있는가? ② 두 식이요법 간에 차이가 있는가?

이들 질문은 통계적 가설 검정의 문제로 해결할 수 있으며 간단한 개념적으로 설명하기로 하자. 먼저 ①과 ②의 질문에 대한 통계적 귀무가설로 [정리 11.1]과 같이 고려할 수 있으며 이는 분할표의 범주에 대한 독립성과 동질성을 검정하는 두 종류의 가설 검정으로 연결된다.

일반적으로 [정리 11.1]의 귀무가설에 언급되는 내용이 적합한지를 알아보는 척도로 피어슨(Karl Pearson: 1857-1936)이 처음 제안한 카이제곱 통계량(X^2 statistic)이 있다.

■ 분할표에 대한 두 종류의 귀무가설

① H_0 : 먹는 약의 종류의 선호도와 나이는 관련이 없다(독립적이다).

② H_0 : 두 식이요법은 차이가 없다(동질적이다).

실제 [보기 11.1]과 [보기 11.2]의 분할표를 통하여 카이제곱 통계량을 구하는 절차를 설명하여 보자. 분할표에서 각 범주에 속하는 칸의 도수는 실제로 관찰된 값으로 칸 관찰도수라 하며 i번째 행과 j번째 열의 칸 (i,j)의 관찰도수를 $O_{ij},\ i=1,\ldots,r; j=1,\ldots,c$라 하자. 특히, i번째 행합계를 $O_{i+}=\sum_{j=1}^{c}O_{ij}$, j번째 열합계를 $O_{+j}=\sum_{i=1}^{r}O_{ij}$로 표시하고 총합계는 $N=\sum_{i=1}^{r}\sum_{j=1}^{c}O_{ij}$라 하자. 관찰도수 O_{ij}에 대응하여 [정리 11.1]의 귀무가설 하에서 칸 (i,j)의 기대도수 E_{ij}는 다음과 같이 각각 구할 수 있다.

$$E_{ij} = \frac{i\text{번째 행합계}\times j\text{번째 열합계}}{\text{총합계}} = \frac{O_{i+}\times O_{+j}}{N} \tag{11.1}$$

따라서 각 칸의 관찰도수와 기대도수와 비교하여 그 차이를 양적으로 평가하는 것이 분할표에서 주된 관심이다. 이를 위한 유용한 통계량이 식 (11.2)의 피어슨의 카이제곱 통계량이다.

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(\text{관찰도수}-\text{기대도수})^2}{\text{기대도수}} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \tag{11.2}$$

보기 11.3

[보기 11.1]의 고객들의 나이와 먹는 약의 종류라는 2개의 범주형 변수의 분할표로부터 관찰도수와 괄호 속의 기대도수를 같이 정리하였다. 여기서 기대도수와 카이제곱 통계량을 구하여 보라.

나 이	먹는 약의 종류		합 계
	캡슐형	정제형	
20세 이상 30세 미만	38	79	117
30세 이상 50세 미만	87	118	205
50세 이상	78	89	167
합 계	203	286	489

1) 분할표에 따르면 총합계는 $N=489$이고 첫 번째 행범주와 열범주에서 20세 이상 30세 미만의 캡슐형 약의 칸 (1, 1)의 관찰도수는 38이고 이에 따른 행과 열의 합계는 다음과 같다.

첫 번째 행합계 : $O_{1+} = 38+79=117$

첫 번째 열합계 : $O_{+1} = 38+87+78=203$

이때 칸 (1, 1)의 기대도수는 다음과 같다.

$$E_{11} = = \frac{O_{1+}\times O_{+1}}{N} = \frac{117\times 203}{489} = 48.6$$

다음으로 첫 번째 행범주와 두 번째 열범주의 경우 칸 (1, 2)의 기대도수와 나머지 칸 (2, 1), (2, 2), (3, 1), (3, 2)의 기대도수는 다음과 같다.

$$E_{12}=\frac{117\times286}{489}=68.4,\ E_{21}=\frac{205\times203}{489}=85.1,\ E_{22}=\frac{205\times286}{489}=119.9,$$

$$E_{31}=\frac{167\times203}{489}=69.3,\quad E_{32}=\frac{167\times286}{489}=97.7$$

2) $$X^2=\sum_{i=1}^{3}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

$$=\frac{(38-48.6)^2}{48.6}+\frac{(79-68.4)^2}{68.4}+.....+\frac{(89-97.7)^2}{97.7}=5.86$$

보기 11.4

[보기 11.2]에서 두 가지 식이요법과 환자의 건강상태라는 2개의 범주형 변수에 따른 분할표로부터 괄호 속의 기대도수와 카이제곱 통계량을 구하여 보라.

식이요법	환자의 건강상태			합 계
	양 호	보 통	불 량	
A	37	24	19	80
B	17	33	20	70
합　계	54	57	39	150

1) 분할표에서 총합계는 $N=150$이고 식이요법 A와 환자의 건강상태가 양호인 첫 번째 행범주와 열범주에서 칸 (1, 1)의 관찰도수는 37이고 이에 따른 행과 열의 합계는 다음과 같다.

첫 번째 행합계 : $O_{1+}=37+24+19=80$

첫 번째 열합계 : $O_{+1}=37+17=54$

이때 칸 (1, 1)의 기대도수는 다음과 같다.

$$E_{11} = \frac{O_{1+} \times O_{+1}}{N} = \frac{80 \times 54}{150} = 28.8$$

다음으로 식이요법 A와 환자의 건강상태가 보통인 칸 (1, 2)의 기대도수와 나머지 칸 (1, 3), (2, 1), (2, 2), (2, 3)의 기대도수는 다음과 같다.

$$E_{12} = \frac{80 \times 57}{150} = 30.4,\ E_{13} = \frac{80 \times 39}{150} = 20.8,\ E_{21} = \frac{70 \times 54}{150} = 25.2,$$

$$E_{22} = \frac{70 \times 57}{150} = 26.6,\ E_{23} = \frac{70 \times 39}{150} = 18.2$$

2) $$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

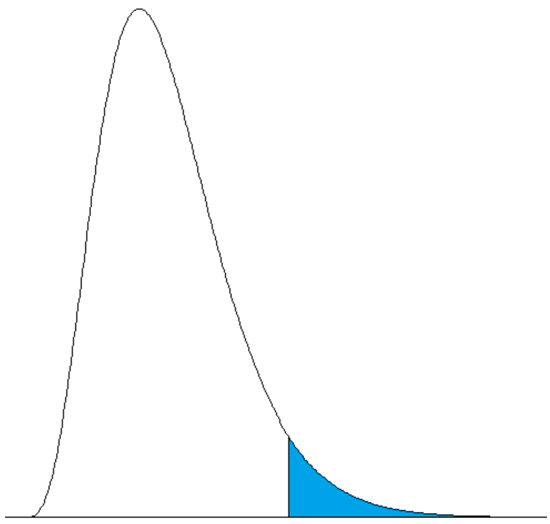
$$= \frac{(37 - 28.8)^2}{28.8} + \frac{(24 - 30.4)^2}{30.4} + \dots + \frac{(20 - 18.2)^2}{18.2} = 8.22$$

11.3 카이제곱 검정

11.2절에서 계산된 카이제곱 통계량 X^2을 이용하여 [정리 11.1]의 귀무가설에 언급되는 내용이 적합한지를 알아보는 절차는 카이제곱 검정(X^2-test)이라는 통계적 기법에 의해서 이루어진다. 특히 [정리 11.1]의 귀무가설 ①과 ②에 대한 카이제곱 검정을 각각 독립성 검정(test of independence)과 동질성 검정(test of homogeneity)이라 한다. 그 이유는 [정리 11.1]의 귀무가설 <① H_0: 먹는 약의 종류의 선호도와 나이는 관련이 없다>는 행범주와 열범주의 관련성을 따지므로 귀무가설의 내용상 <관련이 없다>는 <서로 독립이다>라고 말하는 데 근거하고 있다. 이런 맥락에서 귀무가설 <② H_0:두 식이요법은 차이가 없다>는 행범주 간의 차이를 따지므로 <차이가 없다>는 <서로 동질이다>라고 말할 수 있다. 어느 검정이든지 귀무가설의 기각은 분할표로부터 계산된 카이제곱 통계량 X^2의 값이 매우 크면 귀무가설을 기각하게 된다.

여기서 카이제곱 통계량 X^2의 값이 어느 정도이면 통계적으로 크다고 할 수 있는지를 알아보기 위해선 카이제곱 통계량 X^2의 분포와 이에 관련된 카이제곱 검정법에 대해서 알아보아야 한다.

일반적으로 $r \times c$ 분할표로부터 계산된 카이제곱 통계량 X^2은 귀무가설이 옳다는 가정 하에 각 칸 기대도수가 대략 5 이상이고 총합계가 큰 경우에 근사적으로 X^2-분포를 따르고 자유도(degree of freedom: df)는 $(r-1)\times(c-1)$ = (행범주의 수-1) ×(열범주의 수-1)이 된다. [그림 11.1]에는 자유도가 10인 X^2-분포를 보이고 있는 데 유의수준 α의 기각치(critical value) $X_{\alpha}^2(10)$도 표시되어 있다. 이는 부록의 통계분포표로부터 쉽게 구해진다.



[그림 11.1] 자유도가 10인 X^2-분포

따라서 유의수준 α에서 카이제곱 통계량 X^2의 관측값 χ^2가 기각치 $X_{\alpha}^{2}(df)$에 의한 기각역

$$\chi^2 \geqq X_{\alpha}^{2}(df) \tag{11.3}$$

을 만족하면 통계적으로 유의하게 귀무가설을 기각하는 카이제곱 검정법을 생각할 수 있다. 결론적으로 [보기 11.1]과 [보기 11.2]의 분할표에 대한 [정리 11.1]의 귀무가설은 [보기 11.3]과 [보기 11.4]에서 계산된 카이제곱 통계량 X^2의 관측값 χ^2가 식 (11.3)의 기각역에 의한 카이제곱 검정법에 따라 유의수준 0.05에서 다음과 같이 요약된다.

보기 11.5

[보기 11.3]과 [보기 11.4]에 대해 카이제곱 통계량에 의한 검정을 시행하시오.

1) 나이별 약의 종류에 따른 선호도 조사:
   - 자유도 : (3-1)(2-1) = 2
   - 유의수준 0.05에서 검정통계량의 관측값 $\chi^2 = 5.86$이 기각치 $X^2_{0.05}(2) = 5.99$에 의한 기각역을 만족하지 하지 않으므로, 귀무가설 H_o를 기각할 수 없다. 따라서 나이와 약의 종류의 선호도는 관련이 없다.

2) 식이요법에 따른 환자의 건강상태:
   - 자유도 : (2-1)(3-1) = 2
   - 유의수준 0.05에서 검정통계량의 관측값 $\chi^2 = 8.22$이 기각치 $X^2_{0.05}(2) = 5.99$에 의한 기각역을 만족하므로, 귀무가설 H_o를 기각할 수 있다. 이는 식이요법 간에 효과의 차이가 있음을 말한다.

추가로 R-프로그램에서는 검정방법으로 검정통계량의 관측값 χ^2과 기각치 $X^2_\alpha(df)$에 의해 검정하는 것 대신에 유의확률 $p-$값$=P[X^2 > \chi^2]$을 제공하여 이 값이 유의수준 α보다 작으면 귀무가설을 기각하는 검정방법을 택하고 있다. 이는 11.4절에서 사용의 예를 볼 수 있다.

■ $r \times c$ 분할표의 행과 열의 독립성과 동질성 검정

- $X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(\text{관찰도수}-\text{기대도수})^2}{\text{기대도수}} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$,
- 자유도: $df = (r-1)(c-1)$
- $\chi^2 \geqq X^2_\alpha(df) \Rightarrow$ 귀무가설 <H_0 : 서로 독립(서로 동질)>을 유의수준 α에서 기각한다.

11.4 R-프로그램

[프로그램 11.1]은 [보기 11.1]의 고객들의 나이와 먹는 약의 종류 간에 독립성 검정을 위한 chisq.test( ) 함수를 활용한 R-프로그램을 보여주고 있다. [결과 11.1]에서 <X-squared = 5.8608>는 카이제곱 통계량으로 자유도가 <df = 2>임을 나타내고 있다. 결과에 대한 해석은 [보기 11.5]를 참고로 하면 되나 특히, R-프로그램에서는 기각치와 기각역을 활용한 검정을 제공하는 대신에 $p-$값을 제공하고 있다. <p-value = 0.05338>은 $p-$값으로 이는 $P[X^2 \geq 5.8608]$에 해당한다. 그리고 유의수준 $\alpha = 0.05$보다도 큰 경우를 나타내고 있어 귀무가설 <H_0: 먹는 약의 종류의 선호도와 나이는 관련이 없다>를 기각하지 못함을 보여주고 있다.

[프로그램 11.1] [보기 11.1] 고객들의 나이와 약의 종류에 대한 독립성 검정

```
row1 = c(38,79)
row2 = c(87,118)
row3 = c(78,89)
data.table = rbind(row1,row2,row3)
chisq.test(data.table)
```

[결과 11.1] [프로그램 11.1]의 수행 결과

```
        Pearson's Chi-squared test

data:  data.table
X-squared = 5.8608, df = 2, p-value = 0.05338
```

[프로그램 11.2]은 [보기 11.2]의 두 가지 식이요법과 환자의 건강상태의 분할표에서 두 식이요법 간의 동질성 검정을 위한 chisq.test( ) 함수를 활용한 R-프로그램을 보여주고 있다. 결과에 대한 해석은 [보기 11.5]를 참고로 하면 되나 [결과 11.2]의 <X-squared = 8.224, df = 2, p-value = 0.01638>는 자유도가 $df = 2$인 카이제곱 통계량의 관측값 $\chi^2 = 8.224$을 나타내며 <p-value = 0.01638>는 $p-$값으로 유의수준 $\alpha = 0.05$보다도 작은 경우를 나

타내어 귀무가설 <H_0: 두 식이요법 간에는 차이가 없다>를 기각함을 보여준다.

[프로그램 11.2] [보기 11.2] 두 식이요법 간의 동질성 검정

```
row1 = c(37,24, 19)
row2 = c(17, 33, 20)
data.table = rbind(row1,row2)
chisq.test(data.table)
```

[결과 11.2] [프로그램 11.2]의 수행 결과

```
	Pearson's Chi-squared test

data:  data.table
X-squared = 8.224, df = 2, p-value = 0.01638
```

11.5 연습문제

11.1 철도청에서는 철도 이용객들의 여행 거리와 구입한 좌석 등급 사이에 관계가 있는지를 알고자 한다. 승객 200명을 대상으로 조사를 실시한 결과는 아래의 표와 같다.

여행 거리	좌석 등급			합계
	입석/자유석	일반석	특석	
1~100km	21	14	6	41
101~200km	18	16	8	42
201~300km	16	17	15	48
301~400km	12	14	21	47
401~500km	6	6	10	22
합계	73	67	60	200

(1) 귀무가설 H_0을 서술하라.

(2) 이 검정은 독립성 검정인가, 동질성 검정인가?

(3) 자유도를 계산하라.

(4) 각 칸의 기대도수를 구하여라.

(5) 검정통계량을 구하여라.

(6) 유의수준 5%에서 어떠한 결론을 내릴 수 있는가?

11.2 30대 600명을 대상으로 그들의 연봉과 학력이 관계가 있는지 알기 위하여 조사한 결과이다. 괄호 속의 값은 기대도수를 정리한 것이다. 유의수준 5%에서 독립성 검정을 하여라.

연봉	학력				합계
	중졸 이하	고졸	대졸	대학원졸	
3천만 원 미만	15(4.6)	25(9.2)	10(13.8)	5(27.5)	55
3~4천만 원	20(13.3)	40(26.7)	70(40.0)	30(80.0)	160
4~5천만 원	10(10.4)	20(20.8)	40(31.3)	55(62.5)	125
5~6천만 원	5(7.9)	10(15.8)	20(23.8)	60(47.5)	95
6천만 원 이상	0(13.8)	5(27.5)	10(41.3)	150(82.5)	165
합계	50	100	150	300	600

11.3 개인병원 의사 100명과 종합병원 의사 150명에게 주당 근무시간을 조사하였다. 개인병원 의사와 종합병원 의사의 근무시간은 동일한지를 알고자 한다.

의사	근무시간				합계
	20-30	30-40	40-50	50-60	
개인병원	16	40	38	6	100
종합병원	8	44	59	39	150
합계	24	84	97	45	250

(1) 귀무가설 H_0을 서술하라.

(2) 각 칸의 기대도수를 구하여라.

(3) 검정통계량을 구하여라.

(4) 유의수준 5%에서 어떠한 결론을 내릴 수 있는가?

(5) 유의수준을 1%로 수정하면 결론은 어떻게 달라지는가?

11.4 자동차 생산회사는 가족 구성원의 수와 자동차의 크기 사이에 관계가 있는지 알고자 한다. 자동차 소유자 800명을 대상으로 조사한 결과를 정리한 표는 다음과 같으며, 괄호 ( ) 속의 값은 기대도수를 정리한 것이다.

가족 수	자동차 크기				합계
	소형	중형	대형	트럭	
1명	20(13.0)	35(26.8)	40(45.5)	35(44.7)	130
2명	20(22.0)	50(45.4)	70(77.0)	80(75.6)	220
3-4명	20(26.0)	50(53.6)	100(91.0)	90(89.4)	260
5명 이상	20(19.0)	30(39.2)	70(66.5)	70(65.3)	190
합계	80	165	280	275	800

(1) 유의수준 5%에서 독립성 검정을 하여라.

(2) 유의수준을 10%로 수정하면 결론은 어떻게 달라지는가?

11.5 남성과 여성의 아침식사 메뉴의 차이를 알아보고자 한다. 아침식사를 판매하는 유명한 가게에서 무작위로 선택된 남성과 여성이 주문한 아침식사는 다음과 같다. 괄호 ( ) 속의 기대도수를 구하고, 유의수준 10%에서 동질성 검정을 하여라.

성별	아침식사 메뉴				합계
	토스트	팬케이크	와플	오믈렛	
남성	47( )	35( )	28( )	53( )	163
여성	65( )	59( )	55( )	60( )	239
합계	112	94	83	113	402

11.6 여행사들은 신부의 연령에 따라 인기 있는 신혼여행 장소가 달라질 수 있다고 주장한다. 이를 확인하기 위하여 최근 결혼한 신부 280명을 대상으로 신혼여행을 보낸 장소를 조사하였다. 조사 결과, 20-29세는 하와이와 푸껫이 각각 15명, 몰디브가 10명, 유럽이 20명이었다. 30-39세는 4곳 모두 25명씩이고, 40-49세는 하와이와 푸껫이 각각 25명, 몰디브와 유럽이 각각 15명이었다. 50세 이상은 하와이가 20명, 푸껫이 10명, 몰디브와 유럽이 각각 5명이었다.

(1) 조사 결과에 따른 자료의 분할표를 작성하라.
(2) 각 칸의 기대도수를 구하여라.
(3) 유의수준 5%에서 독립성 검정을 하여라.

11.7 다음은 2005년부터 2010년까지 6년간 유럽연합과 미국의 평균 에너지 사용량을 나타낸 표이다. 괄호 ( ) 속의 값은 기대도수를 정리한 것이다. 연도에 따른 두 지역의 에너지 사용량은 같다고 할 수 있는지 유의수준 5%에서 검정하여라.

연도	지역		합계
	유럽연합	미국	
2010	3413(3361.0)	7164(7216.0)	10557
2009	3302(3291.7)	7057(7067.3)	10359
2008	3505(3493.2)	7488(7499.8)	10993
2007	3537(3589.2)	7758(7705.8)	11295
2006	3595(3588.2)	7697(7703.8)	11292
2005	3613(3641.6)	7847(7818.4)	11460
합계	45011	20965	65976

11.8 한 심리학자는 경제학 전공과 사회과학 전공의 성격 유형에 차이가 있는지를 알고자 한다. 자료에 대한 분할표는 다음과 같으며, 괄호 ( ) 속의 값은 기대도수를 정리한 것이다. 유의수준 5%에서 동질성 검정을 하여라.

전공	성격 유형					합계
	개방적	성실함	외향적	쾌활함	신경적	
경제학	41(47.6)	52(53.5)	46(45.9)	61(59.3)	58(51.8)	258
사회과학	72(65.4)	75(73.5)	63(63.1)	80(81.7)	65(71.2)	355
합계	113	127	109	141	123	613

부록 I

확률분포표

R과 함께하는

통계학의 이해

부록 I 확률분포표

유의수준(significance level)에서 각 분포의 기각치를 찾기 위한 확률분포표는 다음의 사이트를 참고하면 된다.

■ 정규분포표

http://www.itl.nist.gov/div898/handbook/eda/section3/eda3671.htm

- $P(Z < z)$: $Z \sim N(0, 1)$
- 예 : 유의수준 5%에서 기각치 $z_{0.05}$ 는 정규분포표에서 0.05에 제일 가까운 0.44950에 해당하는 행의 1.6과 열의 0.04를 더한 1.64가 된다.

■ t-분포표

http://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm

- $t_{1-\alpha}(\nu)$: 자유도 ν, 유의수준 α
- 예 : 유의수준 5%에서 자유도 23인 기각치 $t_{0.05}(23)$은 t-분포표에서 $t_{1-0.05}(23) = t_{0.95}(23) = 1.714$에 해당한다.

■ 카이제곱 분포표

http://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm

- $X^2_{1-\alpha}(\nu)$: 자유도 ν, 유의수준 α
- 예 : 유의수준 5%에서 자유도 2인 기각치 $X^2_{0.05}(2)$는 카이제곱 분포표에서 $X^2_{1-0.05}(2) = X^2_{0.95}(2) = 5.991$에 해당한다.

■ F-분포표

http://www.itl.nist.gov/div898/handbook/eda/section3/eda3673.htm

- $F_\alpha(\nu_1, \nu_2)$: 자유도 $df_1 = \nu_1$, $df_2 = \nu_2$, 유의수준 α
- 예 : 유의수준 5%에서 자유도 (3, 20)인 기각치 $F_{0.05}(3, 20)$은 F-분포표에서 $F_{0.05}(3, 20) = 3.098$에 해당한다.

부록 II

R 설치 및 기본 사용법

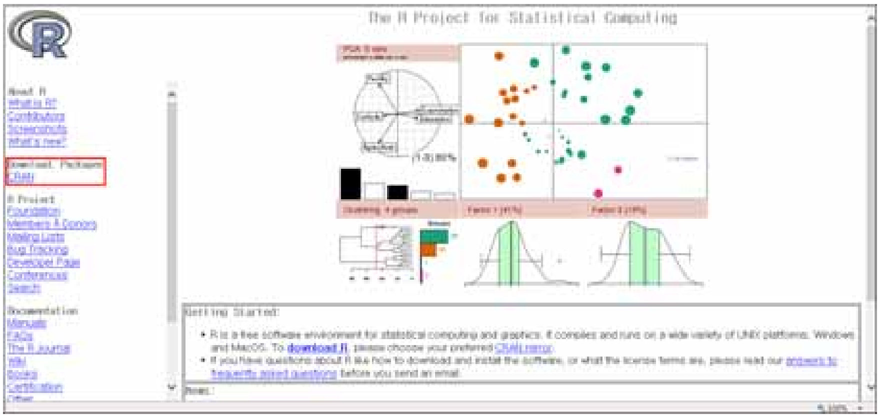
R과 함께하는

통계학의 이해

1. R의 설치

(1) 인터넷에서 [http://www.r-project.org/]로 접속한다.

(2) 화면 왼쪽의 메뉴에서 [Download, Packages]의 [CRAN]을 클릭한다.



(3) 제시된 목록에서 [Korea]를 찾아 [http://cran.nexr.com/]
또는 [http://biostat.cau.ac.kr/CRAN/] 중 하나를 클릭한다.



(4) 다음의 화면에서 [Download R for Windows]를 클릭한다.



(5) 다음의 화면에서 [base]를 클릭한다.

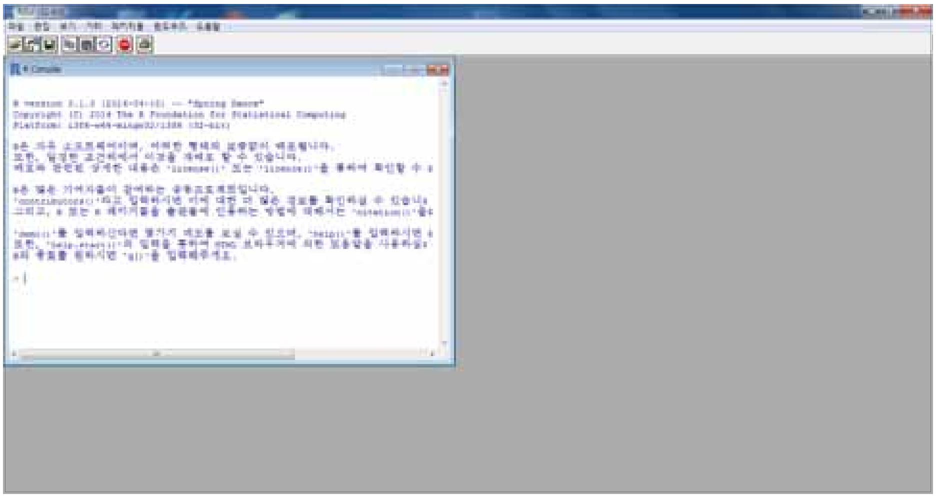


(6) 다음의 화면에서 [Download R 3.1.0 for Windows]를 클릭하고, R 3.1.0 버전의 설치 파일인 [R-3.1.0-win.exe]를 다운로드하여 설치하면 된다.

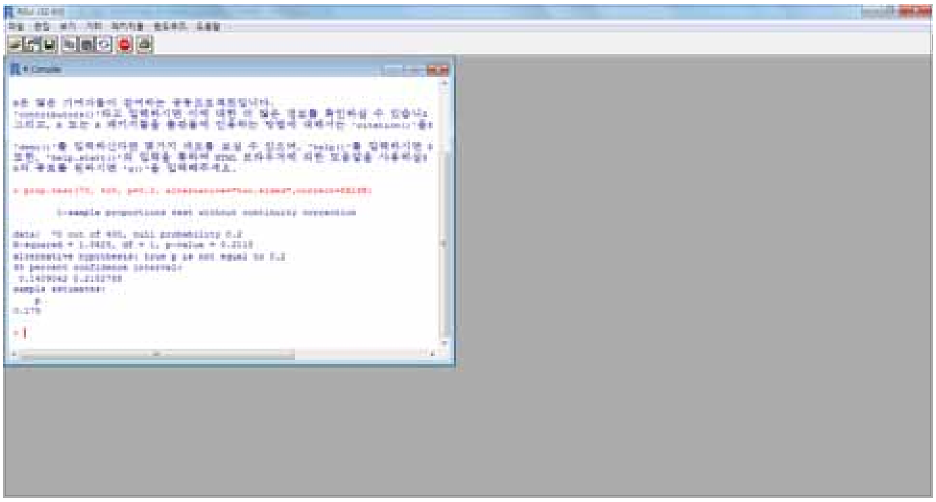


2. R의 기본 사용법

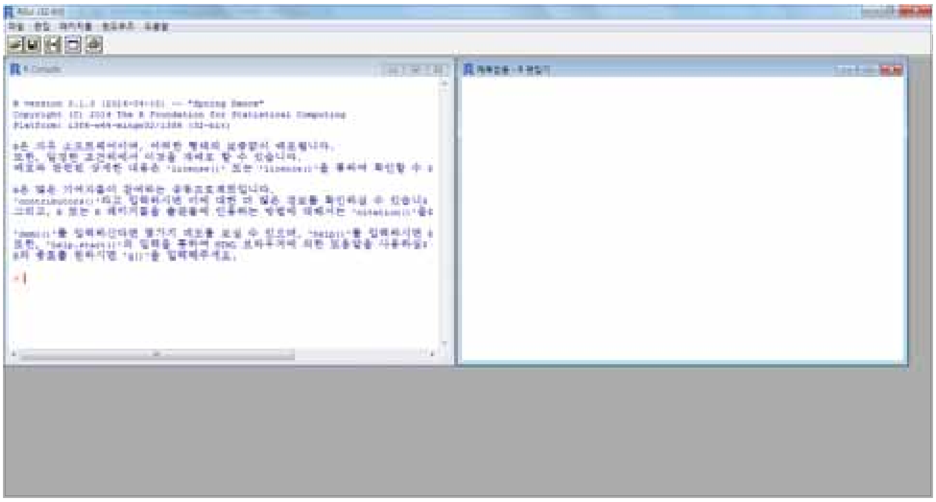
R을 실행하면 다음과 같이 [R Console] 창이 기본적으로 열린다.



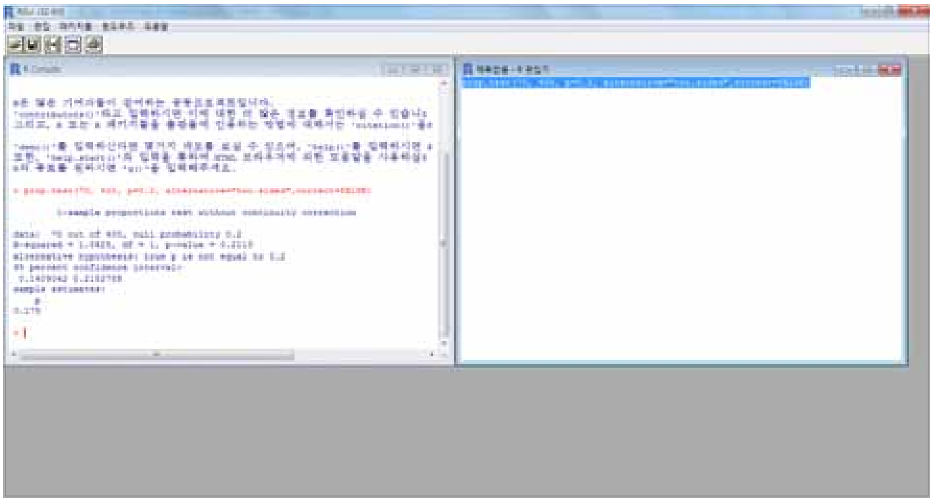
[R Console] 창에 명령문을 입력하고 [Enter]를 누르면 다음과 같이 명령문이 실행되어 그 결과가 출력된다.



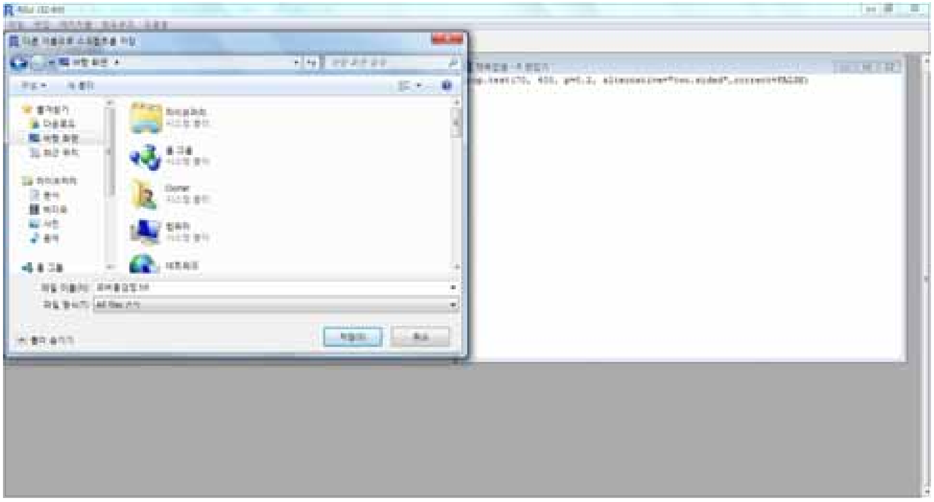
명령문을 입력하는 또 다른 방법은 [파일] - [새 스크립트]를 클릭하면 열리는 [R 편집기]를 활용하는 것이다. 다음의 화면에서 오른편이 [R 편집기] 창이다.



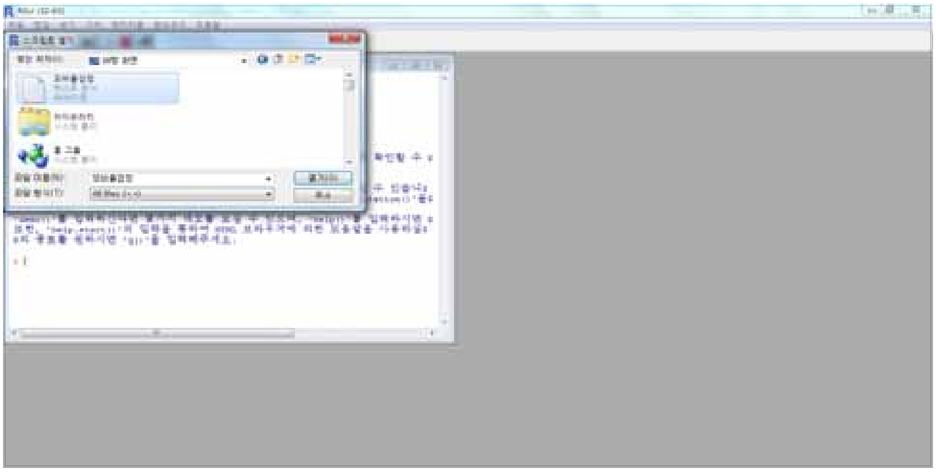
먼저 [R 편집기]에 명령문을 입력한 다음, 실행하고자 하는 명령문을 드래그하여 선택한다. 그 다음 도구모음의 실행 버튼( )을 클릭하거나, [F5] 또는 [Ctrl+R]을 누르면 선택된 명령문이 실행되어 [R Console] 창에 그 결과가 출력된다.



[R 편집기]에 입력된 명령문은 수정 및 편집이 쉽게 가능하고 텍스트 파일로 저장할 수 있어 [R Console] 창보다 편리하다.
[R 편집기]에 입력된 명령문을 다음에 다시 사용하기 위해서는 [파일]-[다른 이름으로 저장]에서 저장하면 된다. 다음과 같이 저장 위치를 지정한 다음, 파일 이름을 [파일명.txt]로 입력하고 파일 유형은 [All files (*.*)]를 선택한다.



이렇게 저장된 파일은 메모장에서 바로 열어볼 수도 있으며, [파일]-[스크립트 열기]에서 다음에 다시 열어 사용할 수 있다.



부록Ⅲ

연습문제 풀이

R과 함께하는

통계학의 이해

1. 통계학의 이해

1.1 (1) 양적 (2) 양적 (3) 질적
(4) 양적 (5) 양적 (6) 질적
(7) 질적 (8) 양적 (9) 양적

1.2 (1) A 피트니스 센터에 등록된 전체 회원들
(2) 등록된 전체 회원들이 일주일에 운동하는 시간의 평균
(3) 등록된 전체 회원들 중 조사에 참여한 회원 50명
(4) 조사에 참여한 회원 50명이 일주일에 운동하는 시간의 평균

1.3 (1) 결혼 상담사가 상담한 모든 고객들
(2) 상담한 모든 고객들 중 결혼을 지속하는 고객들의 비율
(3) 상담한 모든 고객들 중 조사에 참여한 고객 100명
(4) 조사에 참여한 고객 100명 중 결혼을 지속하는 고객들의 비율

1.4 (1) 구간 (2) 명목 (3) 비율
(4) 서수 (5) 명목 (6) 구간
(7) 비율 (8) 서수

1.5 (1) 모든 통계학 수강생들
(2) 통계학 수강생들 중 조사에 참여한 학생 10명
(3) 3.5일

1.6 (1) 연속형 (2) 이산형
(3) 연속형

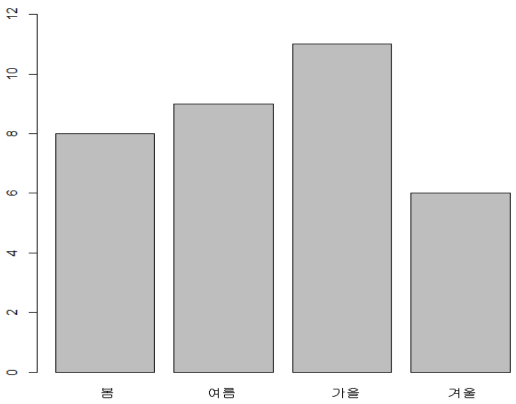
1.7 변환 전: 양적 - 연속형 - 비율척도
변환 후: 질적 - 명목형 또는 순서형

2. 자료의 정리 및 요약

2.1 (1)

계절	도수	상대도수
봄	8	0.235
여름	9	0.265
가을	11	0.324
겨울	6	0.176
합계	34	1.000

(2)



(3) 117 °

2.2 (1) 79.3, 80 (2) 2.44, 2

2.3 (1) 142.68, 11.94

(2) 2.03, 1.42

2.4 (1) 7 (2) 2

2.5 (1) 54

(2)

계급구간(명)	도수	상대도수
11.5이상 65.5미만	19	0.543
65.5이상 119.5미만	9	0.257
119.5이상 173.5미만	1	0.029
173.5이상 227.5미만	4	0.114
227.5이상 281.5미만	2	0.057
합계	35	1.000

2.6 (1) 39 (2) 69

2.7 (1) 1.92 (2) 3.5 (3) 1.5

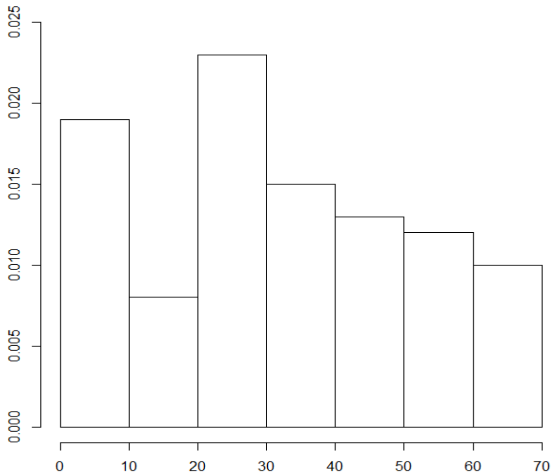
2.8 (1)

계급구간(세)	상대도수
0이상 10미만	0.189
10이상 20미만	0.080
20이상 30미만	0.228
30이상 40미만	0.150
40이상 50미만	0.131
50이상 60미만	0.119
60이상	0.103

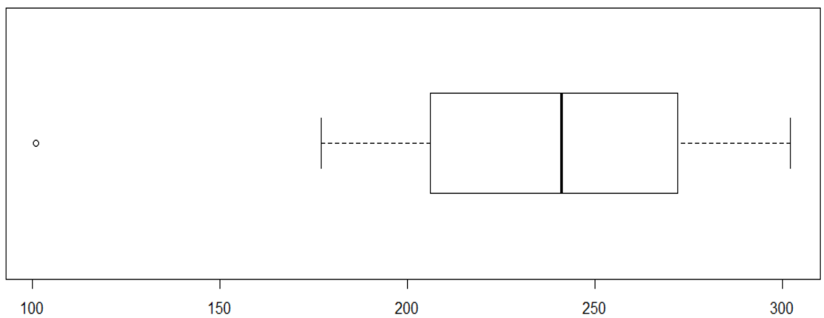
(2)

계급구간(세)	높이
0이상 10미만	0.019
10이상 20미만	0.008
20이상 30미만	0.023
30이상 40미만	0.015
40이상 50미만	0.013
50이상 60미만	0.012
60이상	0.010

(3)



2.9 (1) 234.96 (2) 241
(3) 206, 272 (4) 66
(5) 107, 371 (6) 101
(7)



3. 이산확률변수 및 분포

3.1 (1) A회사:

x	500만원	100만원	-100만원
$P(X=x)$	0.1	0.3	0.6

B회사:

x	300만원	100만원	-100만원
$P(X=x)$	0.2	0.4	0.4

C회사:

x	600만원	0원	-100만원
$P(X=x)$	0.1	0.7	0.2

(2) 20만 원, 60만 원, 40만 원, B회사

(3) C회사, A회사

3.2 (1) 0.2 (2) 2.35, 1.52

(3) 2-3명일 가능성이 크다.

3.3 (1) 1이 나오는 주사위의 개수

(2) 0, 1, 2, 3, 4, 5, 6

(3) $X \sim \text{Bin}(6, 1/6)$

(4) 1, $\sqrt{5/6}$ (5) 0.00002

3.4 (1) 통계학 학사 학위를 얻기 위해 걸리는 기간

(2) 통계학 학사 학위를 얻기 위해서 최소한 3년 이상 걸린다.

(3) 4.85년

3.5 (1) 4.43경기 (2) 0.1476
(3) 0.4734

3.6 (1) 주무기가 포일이 아닌 선수의 수
(2) $X \sim \text{Bin}(25, 0.4)$ (3) 10명
(4) 1.13×10^{-10}

3.7 (1) $X \sim \text{Bin}(3, 1/6)$
(2) -1, 1, 2, 3(만 원)
(3)

x	y(만원)	$P(X=x)$ 또는 $P(Y=y)$
0	-1	0.5787
1	1	0.3472
2	2	0.0694
3	3	0.0046

(4) 0.5, -789원
(5) 0.42, 동일함

3.8 (1) 0.12 (2) 0.11 (3) 0.77
(4) 1.82, 1.4, 서울이 더 많다.

3.9 -614원, 옳지 않다.

3.10 (1) 2.16명 (2) 0.9511
(3) 0.3702

4. 연속확률변수 및 분포

4.1 (1) ○ (2) × (3) ×

4.2 (1) 20 (2) 5.75 (3) 6.5

4.3 (1) 0.9418 (2) 0.3582

4.4 (1) 0.0918 (2) 130.825 이상

4.5 0.9791

4.6 (1) -1.1 (2) 0.3446
(3) 29.25 이상

4.7 (1) 0.0003 (2) 0.0005

4.8 (1) 0.0087 (2) 0.759

5. 표집분포와 중심극한정리

5.1

$\bar{x}$	1	1.5	2	2.5	3	3.5	4
$f(\bar{x})$	1/16	2/16	3/16	4/16	3/16	2/16	1/16

5.2 (1) $\mu = 2.5,\ \sigma^2 = 1.25$

(2) $E(\overline{X}) = 2.5,\ Var(\overline{X}) = 0.625$

(3) 같다.

5.3 $\overline{X} \sim N\left(1910, \dfrac{220^2}{60}\right)$

5.4 0.9033

5.5

$\bar{x}$	3	3.7	4.3	5
$f(\bar{x})$	1/8	3/8	3/8	1/8

5.6 (1) 250, 11.47

(2) $\overline{X} \sim N(250, 11.47^2)$

(3) 0.1922

5.7 (1) $\overline{X} \sim N\left(30, \frac{4^2}{100}\right)$ (2) 0.9876

5.8 0.0129

6. 추정

6.1 (239.84, 248.16)

6.2 (1) 0.791, 0.018

(2) (0.761, 0.821)

6.3 (1) (70.21, 71.79)

(2) (70.15, 71.85)

6.4 (1) 2.511, 0.318

(2) (2.267, 2.755)

6.5 (2.422, 2.600)

6.6 (2.315, 2.707)

6.7 (0.532, 0.668)

6.8 (1.959, 2.041)

6.9 (7.944, 8.456)

6.10 (1) (4.580, 7.420)

(2) (4.268, 7.732)

(3) (3.585, 8.415)

7. 추정

7.1 (1) $H_0:\mu=34$, $H_1:\mu\neq 34$, 양측 (2) $H_0:p=0.29$, $H_1:p\neq 0.29$, 양측

(3) $H_0:\mu=3000$, $H_1:\mu<3000$, 단측 (4) $H_0:p=0.11$, $H_1:p>0.11$, 단측

(5) $H_0:p=0.8$, $H_1:p<0.8$, 단측 (6) $H_0:\mu=6$, $H_1:\mu\neq 6$, 양측

(7) $H_0:\mu=735$, $H_1:\mu>735$, 단측 (8) $H_0:p=0.5$, $H_1:p\neq 0.5$, 양측

7.2 (1) 제1종 오류: 은퇴 전 일하는 평균 기간이 34년일 때, 34년이 아니라고 결론 내리는 경우, 제2종 오류: 은퇴 전 일하는 평균 기간이 34년이 아닐 때, 34년이라고 결론 내리는 경우

(2) 제1종 오류: 고등학교 3학년 학생 중 음주를 하는 학생이 29%일 때, 29%가 아니라고 결론 내리는 경우, 제2종 오류: 고등학교 3학년 학생 중 음주를 하는 학생이 29%가 아닐 때, 29%라고 결론 내리는 경우

(3) 제1종 오류: 대학교 졸업자의 평균 초기 연봉이 3,000만 원보다 많거나 같을 때, 3,000만 원보다 적다고 결론 내리는 경우, 제2종 오류: 대학교 졸업자의 평균 초기 연봉이 3,000만 원보다 적을 때, 3,000만 원보다 많거나 같다고 결론 내리는 경우

(4) 제1종 오류: 여성들의 유방암 발생 가능성이 11%보다 작거나 같을 때, 11% 이상이라고 결론 내리는 경우, 제2종 오류: 여성들의 유방암 발생 가능성이 11% 이상일 때, 11%보다 작거나 같다고 결론 내리는 경우

7.3 (1) $H_0:\mu=7$, $H_1:\mu<7$

(2) $\overline{X}\sim t(21)$ (3) $t=-1.847$

(4) H_0기각 (5) 제1종 오류

7.4 (1) $H_0:p=0.467$, $H_1:p\neq 0.467$

(2) $z=-1.645$ (3) H_0채택

7.5 $z=1.698$, H_0기각

7.6 $p=0.0448$, H_0기각

7.7 $z=3.5$, H_0기각

7.8 (1) $t=1.967$, H_0기각

(2) $p=0.0375$, H_0기각, H_0채택

7.9 (1) $t=1.967$, H_0채택

(2) $p=0.0750$, H_0채택

7.10 $z=-0.2749$, H_0채택

8. 가설 검정

8.1 (1) a (2) e (3) c
(4) b (5) d

8.2 (1) $H_0: p_1 - p_2 = 0$,
$H_1: p_1 - p_2 < 0$
(2) $z = -2.511$ (3) H_0기각

8.3 (1) $H_0: \mu_1 - \mu_2 = 0$,
$H_1: \mu_1 - \mu_2 < 0$
(2) $df = 18$ (3) $t = -2.457$
(4) H_0기각

8.4 (1) $H_0: \mu_1 - \mu_2 = 0$,
$H_1: \mu_1 - \mu_2 < 0$
(2) $t = -6.095$ (3) H_0기각

8.5 (1) 평균 (2) 양측
(3) $H_0: \mu_1 - \mu_2 = 0$,
$H_1: \mu_1 - \mu_2 \neq 0$
(4) 정규분포 (5) $z = 5.421$
(6) H_0기각

8.6 (1) $H_0: \mu_D = 0$, $H_1: \mu_D < 0$
(2) $t = -338.124$ (3) H_0기각

8.7 (1) $H_0: p_1 - p_2 = 0$,
$H_1: p_1 - p_2 \neq 0$
(2) $z = -2.950$ (3) H_0기각

8.8 (1) $H_0: \mu_1 - \mu_2 = 0$,
$H_1: \mu_1 - \mu_2 \neq 0$
(2) $z = -2.735$ (3) H_0채택
(4) H_0기각

8.9 (1) $H_0: \mu_1 - \mu_2 = 0$,
$H_1: \mu_1 - \mu_2 < 0$
(2) $z = -0.248$ (3) H_0채택

8.10 (1) $H_0: \mu_1 - \mu_2 = 0$,
$H_1: \mu_1 - \mu_2 \neq 0$
(2) $t = -1.967$
(3) H_0채택
(4) H_0기각

8.11 (1) $H_0: \mu_D = 0$, $H_1: \mu_D > 0$
(2) $t = 1.754$ (3) H_0채택

9. 분산 분석

9.1 (1) $H_0: \mu_1 = \mu_2 = \mu_3$, H_1: 모든 μ_i 가 다 같은 것은 아니다.

(2) $t-1=2$, $n-t=12$

(3) F-분포

(4) $SSt=23.212$, $SSE=208.324$

(5) $F=0.669$ (6) H_0채택

9.2 (1) $F=8.689$, H_0기각

(2) $F=15.279$, H_0기각

9.3 ㉠: 4 ㉡: 10 ㉢: 237.3

㉣: 48.9 ㉤: 2.06

9.4

요인	자유도	제곱합	평균제곱합	분산비
처리	3	13.032	4.344	0.885
오차	15	73.600	4.907	
전체	18	86.632		

9.5 $F=0.639$, H_0채택

9.6 $F=4.220$, H_0채택

요인	자유도	제곱합	평균제곱합	분산비
처리	4	0.903	0.226	4.220
오차	15	0.803	0.054	
전체	19	1.706		

9.7 (1)

요인	자유도	제곱합	평균제곱합	분산비
처리	2	1468909	734454.6	3.126
오차	12	2819077	234923.1	
전체	14	4287986		

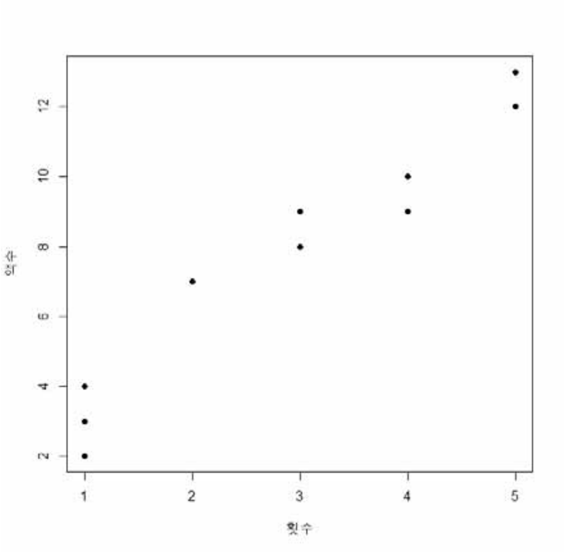
(2) H_0채택

(3) H_0기각

9.8 $F = 4.081$, H_0기각

10. 상관분석과 회귀분석

10.1 (1)



(2) $r = 0.9698$

(3) 높은 양의 상관

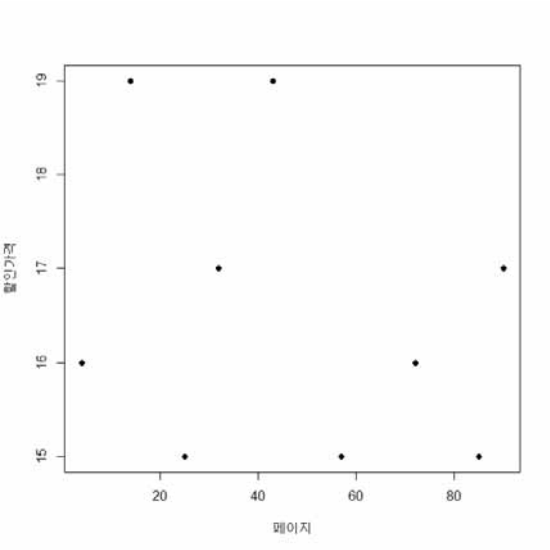
10.2 $\hat{y} = 1.152 + 2.258x$,

$R^2 = 0.9405$

10.3 (1) $\hat{y} = -376.298 + 0.227x$

(2) 80.88세

10.4 (1)



(2) $r = -0.2752$,

약한 음의 상관

(3) $\hat{y} = -0.0141 + 17.2167x$

(4) $SSE = 18.689$, $SSR = 1.531$,

$SST = 20.22$

(5) $R^2 = 0.0757$

(6) $F = 0.5734$, H_0채택

10.5 (1) $\hat{y} = -745.39 + 54.76x$

(2) $F = 19.996$, H_0기각

(3) 1992.61(천원)

10.6 (1) $\hat{y} = 65.086 + 7.095x$

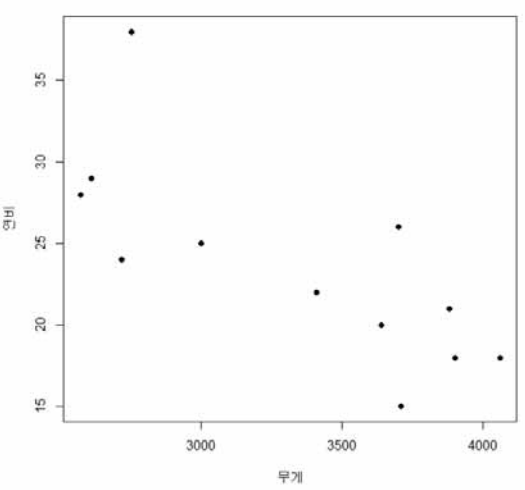
(2) 72.18cm, 143.13cm

(3) 504.98cm, 타당하지 않다.

10.7 (1) $\hat{y} = 35.589 - 0.192x$

(2) $r = -0.5787$, $R^2 = 0.3349$

10.8 $r = -0.7999$



10.9 (1) $\hat{y} = 48.575 - 0.0073x$

(2) $R^2 = 0.6398$

(3) $F = 17.76$, H_0기각

11. 분할표 자료분석

11.1 (1) H_0: 철도 이용객들의 여행 거리와 구입한 좌석 등급은 관계가 없다(독립적이다).

(2) 독립성 검정

(3) $df = 8$

(4) 각 칸의 기대도수는 괄호 ( )속의 값과 같다.

여행 거리	좌석 등급			합계
	입석/자유석	일반석	특석	
1~100km	21(15.0)	14(13.7)	6(12.3)	41
101~200km	18(15.3)	16(14.1)	8(12.6)	42
201~300km	16(17.5)	17(16.1)	15(14.4)	48
301~400km	12(17.2)	14(15.7)	21(14.1)	47
401~500km	6(8.0)	6(7.4)	10(6.6)	22
합계	73	67	60	200

(5) $X^2 = 15.92$

(6) H_0기각

11.2 $X^2 = 255.77$, H_0기각

11.3 (1) H_0: 개인병원 의사와 종합병원 의사의 근무시간은 차이가 없다(동질적이다).

(2) 각 칸의 기대도수는 괄호 ( )속의 값과 같다.

의사	근무시간				합계
	20 - 30	30 - 40	40 - 50	50 - 60	
개인병원	16(9.6)	40(33.6)	38(38.8)	6(18.0)	100
종합병원	8(14.4)	44(50.4)	59(58.2)	39(27.0)	150
합계	24	84	97	45	250

(3) $X^2 = 22.5$

(4) H_0기각

(5) H_0기각

11.4 (1) $X^2 = 15.83$, H_0채택

(2) H_0기각

11.5 $X^2 = 4.01$, H_0채택

11.6 (1)

연령	신혼여행 장소				합계
	하와이	푸켓	몰디브	유럽	
20 - 29세	15	15	10	20	60
30 - 39세	25	25	25	25	100
40 - 49세	25	25	15	15	80
50세 이상	20	10	5	5	40
합계	85	75	55	65	280

(2) 각 칸의 기대도수는 괄호 ( )속의 값과 같다.

연령	신혼여행 장소				합계
	하와이	푸켓	몰디브	유럽	
20 - 29세	15(18.2)	15(16.1)	10(11.8)	20(13.9)	60
30 - 39세	25(30.4)	25(26.8)	25(19.6)	25(23.2)	100
40 - 49세	25(24.3)	25(21.4)	15(15.7)	15(18.6)	80
50세 이상	20(12.1)	10(10.7)	5(7.9)	5(9.3)	40
합계	85	75	55	65	280

(3) $X^2 = 15.70$, H_0채택

11.7 $X^2 = 2.74$, H_0채택

11.8 $X^2 = 3.01$, H_0채택

색인

찾아보기

R과 함께하는

통계학의 이해

A

B

C

D

E

F

H

I

L

M

최용석

1992년 부산대학교 통계학과 교수가 된 후로 현재 부산대학교 자연과학대학 학장과 통계학과 교수로 다변량통계학, 형상분석, 통계조사, 통계상담, 범주형 자료분석에 관심이 많다. 1984년도에 부산대학교 계산통계학과를 졸업하고, 1986년과 1991년에 고려대학교 통계학과 석사와 박사학위를 취득하였다. 1996년 영국 애버딘대학 수리과학부 통계전공 Post-Doc.을 하였다. 2003년에는 영국 리즈대학 방문교수, 2012년에는 미국 아이오와 주립대학 PPSI 연구소 방문교수를 하였다. 호는 일풍(一風: 한 줄기 바람)으로 풍광이 수려한 부산 용호동 섶자리(최초 염전지) 근처에 산다. 개인 블로그(blog.daum.net/musigma)에 개인적 취미 생활인 詩와 그림이 있다. 저자의 논문, 저서 그리고 제자 등 보다 자세한 소개는 홈페이지(yschoi.pusan.ac.kr)를 참고할 수 있겠다.

[빅북] R과 함께하는 통계학의 이해

발행일 2014년 8월 31일
저작권자 빅북운동본부
대표자 조영복
작성자 최용석
주소 부산광역시 금정구 구서2동 248-10 현대빌딩 2F
문의처 051-510-2570 홈페이지 http://bigbook.or.kr/

발행처 교보문고 퍼플
출판등록 2012년 09월 07일 제3-2012-167호
주소 서울시 종로구 종로1가 1번지
대표전화 1544-1900
홈페이지 www.kyobobook.co.kr

편집디자인 좋은땅출판사
홈페이지 www.g-world.co.kr
대표전화 02-374-8616

ISBN 978-89-24-01457-0 (93310)